

UNIVERSIDAD DE DEUSTO

ESTRATEGIAS DE COMUNICACIÓN DE LOS PARTIDOS POLÍTICOS EN SITIOS WEB DE INTERACCIÓN SOCIAL

AGUSTÍN ZUBILLAGA REGO

Bilbao, Agosto de 2015



UNIVERSIDAD DE DEUSTO

ESTRATEGIAS DE COMUNICACIÓN DE LOS PARTIDOS POLÍTICOS EN SITIOS WEB DE INTERACCIÓN SOCIAL

Tesis doctoral presentada por
AGUSTÍN ZUBILLAGA REGO
dentro del Programa de Doctorado en
INGENIERÍA EN INFORMÁTICA Y TELECOMUNICACIÓN

Dirigida por el
Dr. PABLO GARCÍA BRINGAS
y por el
Dr. IKER PASTOR LÓPEZ

Autor | Co-director | Co-director

Bilbao, Agosto de 2015

A la mujer con paciencia y amor infinitos.
Por quien realmente ha merecido la pena todo este esfuerzo.
A Cristina

Resumen

Las últimas décadas han sido testigo de profundos cambios sociales, políticos y tecnológicos. La irrupción de Internet ha cambiado la forma en la que los ciudadanos se relacionan, configurando al mismo tiempo nuevas formas de expresión política.

Dos son las grandes perspectivas respecto a la capacidad de cambio real de la política a través de este nuevo escenario en que la tecnología impregna cada vez más espacios la vida de las personas. El enfoque optimista propone una visión posibilista y transformadora de la tecnología, mientras que la denominada corriente pesimista considera que la vía digital no ofrece sino un nuevo canal en el que se sigue repitiendo el mismo repertorio de acción política tradicional.

En particular, el despegue de tecnologías para la interacción social a través de Internet ha permitido el desarrollo de nuevas capacidades de comunicación entre personas y organizaciones, articulando una nueva esfera social digital. En consecuencia, se abre una oportunidad hacia un modelo potencialmente más democrático y participativo.

Los partidos políticos no se han mantenido ajenos a este vector de cambio, y han incorporado estas novedosas tecnologías que les permiten continuar desarrollando su acción persuasiva orientada finalmente hacia la obtención del voto de los ciudadanos.

Esta tesis propone un nuevo enfoque para interpretar la acción comunicativa de los partidos políticos en España a través de la red social Facebook en el contexto de una contienda electoral. Y lo hace mediante un modelo de comportamiento construido a través de métodos de recuperación de información, minería de opiniones y técnicas de aprendizaje automático supervisado.

Este modelo se fundamenta en tres pilares centrales: (i) la estrategia de comunicación que desarrollan los partidos a través de marcos

comunicacionales que, usados frecuentemente en la comunicación política, permiten delimitar la construcción mental y social de un determinado asunto;(ii) la dimensión deliberativa de las discusiones políticas que se establecen en estas interacciones sociales, y que está intrínsecamente vinculado a su carácter democrático; (iii) y por último, desde el análisis funcional, el modelo es capaz de reconocer la función estratégica de las conversaciones políticas que desempeña cada una de las interacciones individuales que forman parte de cada conversación política.

Abstract

Recent decades have witnessed profound social, political and technological changes. The emergence of the Internet has changed the way citizens relate each others, besides setting new forms of political expression.

There are two major perspectives facing the ability of real change in politics through a new scenario where technology increasingly pervades life spaces of people. The optimistic approach poses a possibilistic and transforming vision of technology, whereas the so-called pessimistic believes that current digital satellite provides only a new channel that keeps repeating the same repertoire of traditional political action.

In particular, the growth of technologies for social interaction via the Internet has allowed the development of new capabilities of communication between people and organizations, giving birth to a new digital social sphere. Consequently, it opens a opportunity into a potentially more democratic and participatory model.

Political parties have not kept outside this path of change, and have incorporated these new technologies that allow them to continue developing their persuasive action in the end oriented to get the vote of citizens.

This thesis proposes a new approach to interpreting communicative action of political parties in Spain through the social network Facebook in the context of an electoral contest. It does so by means of a behavioral model built through information retrieval methods, opinion mining and supervised machine learning techniques.

This model is based on three main pillars: (i) the communication strategy developed by parties through communicational frames often used in political communication, allow to delimit mental and social construction of a particular case; (ii) the deliberative dimension of the political discussions that are established in these social

interactions, and is intrinsically linked to its democratic character; (iii) and finally, from the functional analysis, the model is able to recognize the strategic role of the political talks which plays each of the individual interactions that are part of every political discussion.

Agradecimientos

Al terminar la regata solo nos acordamos de quien se lleva la bandera. Estas palabras, en cambio, son para quienes han estado remando durante el invierno.

Es momento de agradecer a quienes han tenido un papel protagonista en esta página de mi vida. Suele ser este un espacio completamente injusto en la medida que unos párrafos no describen ni el grado de compromiso ni la profundidad de mi agradecimiento sincero. Por ello, mis disculpas previas.

Cuando en 2009 decidí emprender un nuevo camino, la puerta de Deusto se abrió de la mano de un verdadero amigo y compañero, Pablo García Bringas. Querido Pablo: quién me dijo que después de tanto estudiar acabaría realizando una tesis doctoral... Estas palabras solo representan una parte muy pequeña de mi agradecimiento por tu apoyo y arengas permanentes. Para tí, el más fuerte de los abrazos.

Iker Pastor, muchas horas y muchas conversaciones acompañan el esfuerzo de este trabajo. No solo has sido mi co-director de tesis; para mí, lo más importante es tu calidad profesional, tu rigor ético y tu incalculable valor humano. Eres un caballero, aunque te empeñes en restarle valor al honor. Y ahora, ¿qué hacemos con esta tesis? Pues, sin dudarlo un segundo, toca seguir creciendo.

Iker y Pablo, con vuestra dirección he conseguido aunar mi pasión por la política con mi formación y experiencia profesional tecnológicas. Enormes gracias.

Unas líneas de agradecimiento tienen que ir necesariamente para José Gaviria. Muchas gracias, *Sete*, porque siempre has estado ahí. Siempre al lado. Hasta el último minuto. Esta tesis lleva tu sello y con él, mi profundo agradecimiento.

No me quiero olvidar de mis compañeros de DeustoTech Computing, en especial a Borja Sanz, Igor Santos, Jorge de la Peña, Sergio

Huerta y Mikel Salazar, con quienes he compartido momentos de esfuerzo, dudas, avances y retrocesos. A todo el laboratorio, un cariñoso abrazo.

Me gustaría agradecer a la Universidad de Deusto y a su Facultad de Ingeniería, lugar donde me he formado como profesional y donde he fortalecido mis valores. No puedo estar más de acuerdo con el lema deustense; nunca fue tan cierto: *Sapientia melior auro*. En especial, un agradecido recuerdo a Asier Perallos, Guillermo Dorronsoro y Jon Sagastagoitia, de quienes he recibido sincero y valioso apoyo durante los últimos años.

Mi más profunda gratitud a mis padres, Rosa Mari y Agustín, que habeis dado todo por mí y por mis hermanos. Sin pedir nada a cambio. Soy lo que soy gracias a vosotros, de quienes he recibido lo más valioso: mis valores. Arantza, Sergio y Gema: ojalá lleguemos a la altura de papa y mama. Marta, Paula, Álvaro, Andrea e Iñigo, un abrazo por vuestra sonrisa que lo inunda todo cada minuto que estoy con vosotros. A todos, siento defraudaros de nuevo; creo que éste tampoco será el último escalón que suba para seguir formándome.

Por último, y en *primer* lugar, Cristina, nadie como tú sabe lo que ha costado este trabajo. Nadie como tú porque esta tesis es el resultado de tu apoyo ilimitado e incondicional. Tantos madrugones para el estudio, trabajo, lectura y escritura. Tantos fines de semana sin salir y a la sombra de un flexo, vacaciones que no fueron, paseos que no dimos, momentos especiales que cediste con humilde generosidad a esta tesis. Contigo he aprendido que cada vez que alguien me pone una piedra en el camino, me ha de servir para subirme sobre ella para llegar cada vez más arriba.

Índice general

Resumen **iii**

Agradecimientos **vii**

Índice general **ix**

Índice de figuras **xv**

Índice de tablas **xix**

1 Introducción **1**

1.1 Los partidos políticos ante el nuevo escenario digital 3

1.2 Un nuevo lugar de encuentro: los sitios web de interacción social 4

1.2.1 Redes en la sociedad 4

1.2.2 Relacionándose socialmente en *Internet*: *Social network sites* o sitios de interacción social 6

1.2.3 Superando la dicotomía real-virtual: *social informatics* . . . 9

1.3 Motivación 10

1.4 Reto general 12

1.5 Hipótesis y objetivos 13

1.6 Metodología de la investigación 15

1.6.1 Una investigación experimental 17

1.6.2 Consideraciones de seguridad y privacidad 18

1.7 Diseño general 18

1.8 Estructura de la tesis 20

2 Revisión de la literatura **23**

2.1 Internet y democracia 2.0: entre la emancipación democrática y la replicación de espacios 25

2.1.1 Ciberpolítica, redes sociales y nuevas movilizaciones 26

2.1.2 Los partidos políticos en Internet: usos de la web 32

2.1.3 La información política en España en la era digital 34

2.1.4 Deliberación democrática y conversaciones *online* 38

2.1.4.1 La estructura de las redes de discusión: un modelo teórico 39

2.2 Estrategia y competición electoral 41

2.2.1 Los ciudadanos ante el voto: perspectivas y modelos de comportamiento electoral 42

2.2.2 Las campañas en el proceso electoral 50

2.3 Comunicación para la competición electoral 54

2.3.1 De la comunicación de masas a la comunicación interpersonal en redes 57

2.3.1.1 Campañas en *Internet* 60

2.3.1.2 Deliberando a través de una comunicación bidireccional 66

2.3.2 Facebook como canal de comunicación política 67

2.3.2.1 Los mensajes en las redes 70

2.3.2.2 Exponiendo las diferencias políticas 71

2.3.3 La comunicación política desde el análisis de marcos 73

2.3.3.1 El enfoque del encuadre de medios 74

2.3.3.2 Tipos de marcos y el proceso de encuadre 78

2.3.3.3 El enmarcado negativo 80

2.3.3.4 La ética como marco 81

2.3.3.5 El proceso de alineamiento de marcos 83

2.4 Métodos de procesado de datos y textos 85

2.4.1 El descubrimiento de conocimiento 85

2.4.2 La minería de datos 86

2.4.3 La minería de textos 87

2.5 Áreas de investigación relacionadas 89

2.5.0.1 La recuperación de información 89
2.5.0.2 El procesamiento del lenguaje natural 90
2.5.0.3 La extracción de información 90
2.6 Métodos para el filtrado de contenidos y textos 91
2.6.1 El filtrado de contenidos 91
2.6.1.1 Enfoques de filtrado web y valoración 93
2.6.2 El filtrado de textos 95
2.7 Sumario 97

3 Modelado y procesado inicial de los datos 99
3.1 Arquitectura de la solución propuesta 100
3.2 Sitio de interaccción social: *Facebook* 100
3.2.1 Comunicación política entre los miembros de Facebook . . 102
3.3 Adquisición y modelos de datos 104
3.3.1 Conjunto de datos de mensajes políticos 105
3.3.2 Conjunto de datos de conversaciones políticas 108
3.4 Etiquetado de los conjuntos de datos 110
3.4.1 Conjunto de mensajes en muros de partidos políticos . . . 111
3.4.1.1 Clasificaciones del análisis funcional 111
3.4.1.2 Clasificaciones del nivel deliberativo 113
3.4.1.3 Clasificaciones del análisis argumentativo 117
3.4.1.4 Clasificaciones de la estrategia comunicativa . . . 120
3.4.2 Conjunto de conversaciones en muros de partidos políticos 125
3.4.2.1 Clasificaciones del análisis funcional 126
3.4.2.2 Clasificaciones del análisis argumentativo 127
3.4.2.3 Clasificaciones del análisis de marcos 128
3.4.3 Discusión de los resultados del etiquetado 129
3.4.3.1 Mensajes políticos 129
3.4.3.2 Conversaciones políticas 137
3.5 Adaptación de los datos 142
3.5.1 Características extraídas de los mensajes 142
3.5.1.1 Propiedades estadísticas 143
3.5.1.2 Propiedades lingüísticas 145

3.5.2 Características extraídas de las conversaciones 146
3.5.2.1 Propiedades generales 146
3.5.2.2 Propiedades estadísticas 146
3.5.2.3 Propiedades comunicacionales de la conversación 147
3.5.2.4 Propiedades del post inicial 147
3.5.2.5 Propiedades del flujo de mensajes de respuesta . 148
3.5.3 Discusión de los resultados 150
3.6 Sumario 155

4 Categorización automática de mensajes y conversaciones políticas 157
4.1 El aprendizaje automático 159
4.1.1 Las redes bayesianas 160
4.1.2 Las máquinas de soporte vectorial 162
4.1.3 Los K-vecinos más cercanos 164
4.1.4 Los árboles de decisión 166
4.2 Validación empírica 167
4.2.1 Metodología 167
4.3 Discusión de los resultados 170
4.3.1 Conjunto de mensajes 171
4.3.1.1 Función informativa 171
4.3.1.2 Función movilizadora 171
4.3.1.3 Grado de debate 173
4.3.1.4 Grado de contribución 174
4.3.1.5 Nivel de meta-comunicación 176
4.3.1.6 Nivel de personalización 177
4.3.1.7 Referencialidad de contenidos 177
4.3.1.8 Nivel de campaña negativa 178
4.3.1.9 Marco primario 179
4.3.1.10 Marco secundario 181
4.3.1.11 Resumen 182
4.3.2 Conjunto de conversaciones 183
4.3.2.1 Función informativa 184
4.3.2.2 Función movilizadora 184

4.3.2.3 Nivel de meta- comunicación 186
4.3.2.4 Nivel de personalización 188
4.3.2.5 Nivel de campaña negativa 188
4.3.2.6 Marco primario 189
4.3.2.7 Marco secundario 190
4.3.2.8 Resumen 192
4.4 Sumario 194

5 Mejora del modelo **195**
5.1 Obteniendo relevancias mediante la técnica de Ganancia de información 196
5.2 Nuevo modelo de mensajes. Discusión de resultados 198
5.2.1 Función informativa 198
5.2.2 Función movilizadora 200
5.2.3 Nivel de meta-comunicación 202
5.2.4 Nivel de personalización 204
5.2.5 Grado de contribución 204
5.2.6 Referencialidad de contenidos 206
5.2.7 Grado de debate político 208
5.2.8 Nivel de campaña negativa 210
5.2.9 Marco primario 213
5.2.10 Marco secundario 213
5.2.11 Análisis del modelo de mensajes 216
5.3 Relaciones entre las variables del modelo de acción comunicativa 217
5.3.1 Análisis de la *Ganancia de Información* en las conversaciones políticas 218
5.3.2 Discusión de resultados 222
5.3.2.1 Función informativa 222
5.3.2.2 Función movilizadora 222
5.3.2.3 Nivel de meta-comunicación 224
5.3.2.4 Nivel de personalización 224
5.3.2.5 Nivel de campaña negativa 227
5.3.2.6 Marco primario 229

5.3.2.7 Marco secundario 231
5.3.2.8 Análisis del modelo de conversaciones 231
5.4 Sumario 235

6 Conclusiones **237**
6.1 Síntesis del trabajo realizado 237
6.2 Resumen de contribuciones 238
6.3 Limitaciones 242
6.4 Aplicación a la sociedad 244
6.5 Trabajo futuro 244
6.5.1 Medir la eficacia real de la comunicación 245
6.5.2 Correspondencia con las redes de difusión e influencia política 245
6.5.3 Incorporación de meta-información de los contenidos . . . 246
6.5.4 Incorporación de detección de sentimientos y representación de entidades 246
6.5.5 Reducción del nivel de variabilidad de los resultados de clasificación 247
6.6 Consideraciones finales 247

Bibliografía **249**

Índice de figuras

1.1 Diseño de la investigación 19

2.1 Framework clásico de la minería de textos. 89

3.1 Arquitectura general del sistema 101
3.2 Ejemplo de *muro* digital. 106
3.3 Estructura de un mensaje dentro de una conversación política. . . 107
3.4 Estructura de una conversación en *Facebook* 109
3.5 Estructura de un *post* o mensaje inicial en una conversación. . . . 110
3.6 Ejemplo de mensaje informativo con referencia a contenidos propios. 112
3.7 Ejemplo de mensaje informativo con referencia a contenidos de terceros. 112
3.8 Ejemplo de mensaje movilizador hacia el voto. 113
3.9 Ejemplo de mensaje movilizador hacia otra acción política. 113
3.10 Ejemplo de mensaje de debate de tipo irrelevante. 114
3.11 Ejemplo de mensaje de debate de tipo reafirmativo. 114
3.12 Ejemplo de mensaje de debate de tipo controvertido 114
3.13 Ejemplo de mensaje de debate de tipo polémico 115
3.14 Ejemplo de mensaje de debate de tipo ofensivo o insulto 115
3.15 Ejemplo de varios mensajes sin referencias. 115
3.16 Ejemplo de mensaje que referencia a contenidos de terceros. . . . 116
3.17 Ejemplo de mensaje de tipo autorreferencial. 116
3.18 Ejemplo de mensaje con contenidos referenciados por un medio de comunicación. 117

3.19 Ejemplo de mensaje de personalización hacia el propio candidato. 117
3.20 Ejemplo de mensaje de personalización hacia el propio partido. . 118
3.21 Ejemplo de mensaje de personalización hacia el otro candidato. . 118
3.22 Ejemplo de mensaje de tipo meta-comunicación que usa eslóganes.119
3.23 Ejemplo de mensaje de tipo meta-comunicación tipo propuestas de acción política. 119
3.24 Ejemplo de mensaje de tipo meta-comunicación con cobertura de la campaña. 120
3.25 Ejemplo de mensaje de tipo campaña negativa con mención a partidos adversarios. 120
3.26 Ejemplo de mensaje de tipo campaña negativa con mención a candidatos adversarios. 121
3.27 Ejemplo de mensaje con marco de tipo conflicto. 122
3.28 Ejemplo de mensaje con marco de tipo interés humano. 122
3.29 Ejemplo de mensaje con marco de tipo consecuencias económicas.123
3.30 Ejemplo de mensaje con marco de tipo responsabilidad. 123
3.31 Ejemplo de mensaje con marco de tipo oponente inmoral. 124
3.32 Ejemplo de mensaje con marco de tipo candidato populista. 124
3.33 Ejemplo de mensaje con marco de tipo candidato héroe. 124
3.34 Ejemplo de mensaje con marco de tipo candidato estratega. 125
3.35 Ejemplo de mensaje con marco de tipo candidato víctima. 125
3.36 Histograma con la distribución de mensajes respecto a la función informativa 130
3.37 Histograma con la distribución de mensajes respecto a la función de movilización 131
3.38 Histograma con la distribución de mensajes respecto a la función informativa. 131
3.39 Histograma con la distribución de mensajes respecto al grado de debate político. 132
3.40 Histograma con la distribución de mensajes respecto al nivel de meta-comunicación. 133
3.41 Histograma con la distribución de mensajes respecto a la referencialidad de contenidos. 134

3.42 Histograma con la distribución de mensajes respecto a la personalización. 135
3.43 Histograma con la distribución de mensajes respecto a la clase de negatividad en la campaña. 135
3.44 Histograma con la distribución de mensajes respecto al marco primario. 136
3.45 Histograma con la distribución de mensajes respecto al marco secundario. 137
3.46 Histograma con la distribución de conversaciones políticas respecto a la función informativa 138
3.47 Histograma con la distribución de conversaciones respecto a la función movilizadora 139
3.48 Histograma con la distribución de conversaciones respecto al nivel de meta-comunicación. 140
3.49 Histograma con la distribución de conversaciones respecto al nivel de campaña negativa 140
3.50 Histograma con la distribución de conversaciones respecto a la clase de marco primario. 141
3.51 Histograma con la distribución de conversaciones respecto a la clase de marco secundario. 142

4.1 Ejemplo de clasificación mediante una red bayesiana 161
4.2 Ejemplo de clasificación con una máquina de soporte vectorial o *SVM* 163
4.3 Ejemplo de clasificación mediante K-vecinos más cercanos 165
4.4 Ejemplo de clasificación mediante un árbol de decisión 167

Índice de tablas

2.1 Cuadro síntesis de las visiones optimistas-pesimistas sobre Internet. 29

3.1 Dimensiones de los mensajes y sus clasificaciones. 111
3.2 Dimensiones de las conversaciones y sus clasificaciones. 126
3.3 Distribución de mensajes con respecto a la función informativa . . 130
3.4 Distribución de mensajes con respecto a la función movilizadora . 130
3.5 Distribución de mensajes con respecto al grado de contribución . 131
3.6 Distribución de mensajes con respecto al grado de debate político 132
3.7 Distribución de mensajes según el nivel de meta-comunicación . . 133
3.8 Distribución de mensajes según la referencialidad de contenidos . 133
3.9 Distribución de mensajes según la personalización 134
3.10 Distribución de mensajes según el nivel de campaña negativa. . . 135
3.11 Distribución de mensajes según el marco primario, incluyendo el desglose por marco de responsabilidad. 136
3.12 Distribución de mensajes según el marco secundario, incluyendo el desglose por marco de responsabilidad. 137
3.13 Distribución de conversaciones con respecto a la función informativa. 138
3.14 Distribución de conversaciones respecto a la función movilizadora.139
3.15 Distribución de conversaciones respecto al nivel de meta-comunicación139
3.16 Distribución de conversaciones respecto al nivel de personalización140
3.17 Distribución de conversaciones respecto al marco primario, incluyendo el desglose por marco de responsabilidad. 141

3.18 Distribución de conversaciones respecto al marco secundario, incluyendo el desglose por marco de responsabilidad. 142

3.19 Número de características para cada conjunto de datos 154

4.1 Resultados de la clasificación de la función informativa en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 172

4.2 Resultados de la clasificación de la función movilizadora en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 173

4.3 Resultados de la clasificación del grado de debate en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 174

4.4 Resultados de la clasificación del grado de contribución en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 175

4.5 Resultados de la clasificación del nivel de meta-comunicación en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 176

4.6 Resultados de la clasificación del nivel de personalización en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 178

4.7 Resultados de la clasificación de la referencialidad de contenidos en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 179

4.8 Resultados de la clasificación del nivel de campaña negativa en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 180

4.9 Resultados de la clasificación del marco primario en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 181

4.10 Resultados de la clasificación del marco secundario en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 182

4.11 Cuadro resumen con las mejores técnicas para categorizar mensajes, en términos de precisión (%) y área bajo la curva ROC (*AUC*).183

4.12 Resultados de la clasificación de la función informativa en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 185

4.13 Resultados de la clasificación de la función movilizadora en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 186

4.14 Resultados de la clasificación del nivel de meta-comunicación en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 187

4.15 Resultados de la clasificación del nivel de personalización en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 189

4.16 Resultados de la clasificación del nivel de campaña negativa en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 190

4.17 Resultados de la clasificación del marco primario en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 191

4.18 Resultados de la clasificación del marco secundario en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC). 192

4.19 Cuadro resumen con las mejores técnicas para clasificación de conversaciones en términos de precisión (%) y área bajo la curva ROC (*AUC*). 193

5.1 Reducción de características (%) en mensajes por aplicación de *Information Gain*. 197

5.2 Reducción de características (%) en conversaciones por aplicación de *Information Gain*. 197

5.3 Resultados comparados del uso de *Information Gain* para clasificar la función informativa en mensajes, en términos de precisión(%) y AUC. 199

5.4 Resultados comparados del uso de *Information Gain* para clasificar la función movilizadora en mensajes, en términos de precisión(%) y AUC. 201

5.5 Resultados comparados del uso de *Information Gain* para clasificar el nivel de meta-comunicación en mensajes, en términos de precisión(%) y AUC. 203

5.6 Resultados comparados del uso de *Information Gain* para clasificar el nivel de personalización en mensajes, en términos de precisión(%) y AUC. 205

5.7 Resultados comparados del uso de *Information Gain* para clasificar el grado de contribución en mensajes, en términos de precisión(%) y AUC. 207

5.8 Resultados comparados del uso de *Information Gain* para clasificar la referencialidad de los contenidos en mensajes, en términos de precisión(%) y AUC. 209

5.9 Resultados comparados del uso de *Information Gain* para clasificar el grado de debate político en mensajes, en términos de precisión(%) y AUC. 211

5.10 Resultados comparados del uso de *Information Gain* para clasificar el nivel de campaña negativa en mensajes, en términos de precisión(%) y AUC. 212

5.11 Resultados comparados del uso de *Information Gain* para clasificar el marco primario en mensajes, en términos de precisión(%) y AUC. 214

5.12 Resultados comparados del uso de *Information Gain* para clasificar el marco secundario en mensajes, en términos de precisión(%) y AUC. 215

5.13 Comparativa de mejores resultados de categorización de mensajes, en términos de precisión (%) y AUC. 217

5.14 Modelo mejorado de clasificación de mensajes, en términos de precisión (%) y AUC. 218

5.15 Análisis de ganancia de información de la función informativa de las conversaciones políticas. 219

5.16 Análisis de ganancia de información de la función movilizadora de las conversaciones políticas. 219

5.17 Análisis de ganancia de información del nivel de personalización de las conversaciones políticas. 220

5.18 Análisis de ganancia de información del nivel de meta-comunicación de las conversaciones políticas. 220

5.19 Análisis de ganancia de información del nivel de campaña negativa de las conversaciones políticas. 220
5.20 Análisis de ganancia de información del marco primario de las conversaciones políticas. 221
5.21 Análisis de ganancia de información del marco secundario de las conversaciones políticas. 221
5.22 Resultados comparados del uso de *Information Gain* para clasificar la función informativa en conversaciones, en términos de precisión(%) y AUC. 223
5.23 Resultados comparados del uso de *Information Gain* para clasificar la función movilizadora en conversaciones, en términos de precisión(%) y AUC. 225
5.24 Resultados comparados del uso de *Information Gain* para clasificar el nivel de meta-comunicación en conversaciones, en términos de precisión(%) y AUC. 226
5.25 Resultados comparados del uso de *Information Gain* para clasificar el nivel de personalización en conversaciones, en términos de precisión(%) y AUC. 228
5.26 Resultados comparados del uso de *Information Gain* para clasificar la campaña negativa en conversaciones, en términos de precisión(%) y AUC. 230
5.27 Resultados comparados del uso de *Information Gain* para clasificar el marco primario en conversaciones, en términos de precisión(%) y AUC. 232
5.28 Resultados comparados del uso de *Information Gain* para clasificar el marco secundario en conversaciones, en términos de precisión(%) y AUC. 233
5.29 Modelo mejorado de clasificación de conversaciones políticas, en términos de precisión (%) y AUC. 234
5.30 Comparativa de mejores resultados de categorización de conversaciones, en términos de precisión (%) y AUC. 234

«No sabemos ni un
cienmillonésimo de nada.»

Thomas Alva Edison (1847-1931)

CAPÍTULO

1

Introducción

DESDE los primeros mítines de los representantes sindicales durante el siglo XIX hasta la primera campaña de Barack Obama en las elecciones presidenciales de Estados Unidos, mucho ha cambiado cómo los actores políticos se comunican. El primer debate televisado de candidatos en Estados Unidos, entre Nixon y Kennedy o la convocatoria de manifestaciones a través de SMS[1] tras el 11-M de 2004 en España son algunos ejemplos de cómo se ha ido produciendo esta transformación. Igualmente, el conjunto de procesos de la primavera árabe en la segunda década del siglo XXI ha mostrado cómo es posible provocar y conducir grandes movilizaciones sociales a través de nuevos instrumentos basados en la tecnología. Y es que la comunicación social y la comunicación política han evolucionado en paralelo al avance de nuevas capacidades, conocimiento y la implantación de tecnologías.

Este cambio ha sido más acentuado durante los últimos años y especialmente las dos últimas décadas. Tras apenas veinte años, millones de personas disponen de conexión a *Internet*, lo que ha hecho considerar *Internet* como el sexto continente [PP09]. Su uso ha pasado de ser esporádico a convertirse en una herramienta cotidiana. Esta enorme Red denominada *Internet*, ha dejado de ser una herramienta con propósito en sí misma, para convertirse en un medio a través del cual realizar cada vez más actividades de la vida diaria. A principios de

[1] SMS: *Short Message Service* es un servicio envío de mensajes cortos desarrollado sobre tecnologías de tecomunicación móvil.

2015, en España 20,6 millones de personas disponen de acceso a *Internet*, más de dos tercera parte tiene cuenta en redes sociales[1] (en una o varias de ellas) e *Internet*, y crecientemente a través de dispositivos móviles inteligentes [Tel15] o *smarphones*[2].

Los usos de esta tecnología constituyen otro vector de cambio. Las personas hemos pasado de actuar como meros lectores de los contenidos que emitían las organizaciones, las empresas o las instituciones, a redactar, grabar, editar y publicar incluso nuestros propios contenidos. Esta Web 2.0[O'R04] ofrece un nuevo escenario en el que casi todo está por descubrir, crear o inventar. La enorme cantidad de información de todo tipo es el resultado, entre otros motivos, de que los ciudadanos participamos de esa producción de contenidos. En la Red, los internautas[3] han pasado de ser viajeros a ser *webactores*, esto es, protagonistas *de* la Red y *en* la Red.

La disponibilidad masiva en el acceso a *Internet* plantea una supuesta mayor democratización de *Internet*, que corre en paralelo al carácter más democrático de la sociedad [Hin10, Sim08, Sav02], de manera bidireccional. Los nuevos movimientos sociales tratan de dar respuesta a necesidades y demandas nuevas -y también clásicas- en un momento en que existe un cierto debate sobre la legitimidad de algunas instituciones políticas y, en especial, de los partidos políticos en determinadas sociedades occidentales. Todo esto confluye en fenómenos que pertenecen al terreno de la sociología, la psicología, las ciencias de la información y la comunicación, la ciencia política o la tecnología. Los cambios sociales afectan y se ven afectados por el comportamiento individual; y la tecnología es transversal -o facilitadora- en esos cambios.

El resto del capítulo se estructura de la siguiente manera. La sección 1.2 presenta el concepto de sitios web de interacción social y los cambios que éstas han traído consigo. Seguidamente, la sección 1.3 explica el marco que justifica la relevancia de esta tesis. Ante esta situación, el reto que se plantea se describe en la sección 1.4, a partir del cual se establece la hipótesis fundamental y los objetivos de esta tesis, que se presentan en la sección 1.5. Tras ello, la sección 1.6 describe la metodología que guía la investigación para, en la sección 1.7 exponer el diseño general sobre el que llevar a cabo la investigación. Por último, en la sección 1.8 se detalla la estructura global del resto de la tesis.

[1] Es un servicio ofrecido a través de *Internet* que permite que personas se conecten entre sí, crear grupos en los que se puede compartir información entre ellos. Esta conectividad social se hace de manera personal y no anónima.

[2] Los *smartphones* o teléfonos inteligentes permiten desarrollar actividades más complejas que un simple teléfono móvil, por lo que disponen de una capacidad de cálculo mayor, además de niveles de conectividad superiores.

[3] Nombre que reciben las personas que utilizan *Internet*.

1.1 Los partidos políticos ante el nuevo escenario digital

Los partidos políticos se han incorporado a esta transformación digital. Desde hace años cuentan con sitios web desde los que informan de sus políticas, sus propuestas, sus estructuras orgánicas o su acción institucional. Comunican digitalmente e incorporan la tecnología como instrumento para procesar su información interna, como forma de expresión y como canal de información para llegar a sus audiencias. El correo electrónico, los SMS, las aplicaciones móviles para *smartphones*, canales propios en *YouTube*[1] y otras redes sociales forman parte de la cotidianidad de la acción de comunicación de los partidos. Unos partidos políticos siempre conectados con el resto de actores políticos y con la sociedad o, en terminología anglosajona, *always on*.

De este modo, en determinadas sociedades modernas, sus sistemas políticos han dado paso desde un control democrático limitado a determinados momentos en el tiempo, a una suerte de democracia continua; una democracia en directo, en la que los ciudadanos han ido adquiriendo más protagonismo y responsabilidad, y en la que los medios de comunicación son cuestionados cuando no ignorados. En estos sistemas, los ciudadanos demandan más transparencia en la medida en que exigen y disponen de más mecanismos de control de la acción de las instituciones políticas.

Del mismo modo, en este nuevo escenario en el que se da un proceso dialéctico entre lo físico y lo digital -o, si se prefiere, entre lo virtual y lo real-, los partidos tratan de saber cómo operar estratégicamente en el ámbito electoral. Entre estas cuestiones adquiere importancia identificar necesidades y demandas de los ciudadanos y de la sociedad en su conjunto, para incorporarlas a las propuestas electorales y de gobierno. En definitiva, tomar el pulso, sondear y conocer a un electorado al que dar respuesta mediante la escucha en *Internet*, y finalmente medir la eficacia de sus mensajes.

En particular, las campañas electorales restringidas únicamente al período previo al día del voto han dado paso a la campaña permanente, como un flujo continuo de comunicación, no ya unidireccional, sino bidireccional o multidireccional [Nim99]. Y es que los partidos políticos ya no son los únicos actores de la campaña, lo que plantea nuevos retos sobre cómo comunicarse e interactuar con una sociedad más abierta.

[1] YouTube es una red social para compartición de vídeos. `http://www.youtube.com`.

1.2 Un nuevo lugar de encuentro: los sitios web de interacción social

1.2.1 Redes en la sociedad

La capacidad para relacionarse con los demás es una de las principales características del ser humano, ya viva en comunidad o en sociedad. La calidad de las relaciones humanas en gran medida es producto de la comunicación entre ellas. Biológicamente el ser humano está construido para imitar la apariencia de los demás, adoptar en cierta manera su estado de ánimo, e incluso las emociones de los que nos rodean. Tal es así que cuando las personas tienen la libertad para hacer lo que les place, lo normal es que tiendan a imitarse. Las redes sociales sirven para transmitir emociones, comportamiento, motivaciones, actitudes y valores [CF10].

Si en las sociedades tradicionales e industriales el entorno de una persona estaba delimitado geográficamente, en la actuales sociedades post-industriales las tecnologías de la información y la comunicación han permitido que dos o más personas se puedan conectar entre sí instantáneamente desde casi cualquier lugar del mundo. En consecuencia, se reduce la distancia geográfica a una distancia digital que se estrecha por momentos. El *zoom politicom* de Aristóteles [Ari07] ha pasado del *homo sapiens* al *homo dictyoius* u *hombre-red* u hombre en red.

Dado que la sociabilidad es una característica intrínseca de los seres humanos, y la conformación y pertenencia a distintas redes es un proceso natural, desde hace décadas se ha constatado un interés creciente por las redes sociales, su descripción, el estudio de su comportamiento y la dinámica de estas redes [CF10, Put01].

En el ámbito digital, la nueva web no tiene todas sus normas de comportamiento totalmente cerradas y definidas, sino que en el dinamismo alcanza a las propias normas, que se crean y cambian continuamente. Así mismo, el comportamiento dinámico en estas redes es diferente de la arquitectura. Ésta última estudia la infraestructura subyacente de nodos que forman la red, mientras que la primera permite descubrir, conocer, describir y predecir el comportamiento que sucede sobre esa estructura. La dinámica representa los flujos representados por los enlaces, cómo cambian estos enlaces, la diferencia de estos flujos, la distancia cambiante de estas relaciones, los intercambios y las modificaciones que se producen en estas redes.

Dentro de una red, cuando un nodo se conecta con otro, tiende a hacerlo con

aquellos que le resultan más afines a él, más similares. Esta característica, denominada *homofilia*, explica en numerosas ocasiones, tanto la creación de redes sociales como su propia evolución a lo largo del tiempo. El contagio o propagación de fenómenos (información, poder, influencia, intereses, etc.) a través de las redes, lo mismo que las sociedades humanas o incluso las agrupaciones animales, ha sido fuente de estudio durante décadas y aún continúa siéndolo [CF10].

Uno de los conceptos que más eficaz resulta para entender la sociabilidad de los grupos humanos es el capital social, formulado por Putnam [Put01]. Articulado en torno al afecto, la confianza mutua, las normas efectivas del colectivo y las redes sociales que se establecen en un grupo, el capital social es clave a la hora de que los grupos humanos puedan construir una sociedad que coopera y colabora. El capital social, que tiene un efecto positivo sobre la calidad democrática de una sociedad, es uno de los que pone mayor énfasis en el valor de las redes sociales. Abundando, el capital social se sustenta en dos tipos de capitales sociales: el *bonding capital* o capital vínculo y el *bridging capital* o capital puente. El *bonding* se pone de manifiesto dentro de un mismo grupo, mientras que el *bridging* es el necesario para construir relaciones con un grupo o colectivo exterior al propio.

Desde la perspectiva técnica, el estudio de las redes lleva varias décadas de desarrollo, planteando diferentes técnicas para su estudio y análisis [AB02]. En particular, las redes complejas son un tipo de redes (es decir, de nodos conectados) que cumplen unas ciertas condiciones (una cola pesada en la distribución de grado, un alto coeficiente de agrupamiento, una estructura comunitaria a muchas escalas, y evidencias claras de una estructura jerarquizada) [Sol]. El análisis de redes ha tenido un campo fértil desde la propia matemática y la física principalmente, pero con aplicación reciente y cada vez más intensa en campos como la biotecnología o las redes de ordenadores. El interés de las redes complejas reside precisamente en que la mayoría de redes sociales pueden ser consideradas redes complejas. El estudio de las redes abarca desde el marketing a la biología, la medicina o las telecomunicaciones [AB02]. La relevancia de estos estudios descansa sobre la importancia que tienen los enlaces que existen entre los distintos nodos. Comprender sus causas, forma y evolución resultan claves en tanto en cuanto son los enlaces los que crean redes [BF14], por lo que la naturaleza y características de dichos vínculos es valioso para poder comprender los efectos sobre la estructura de las redes.

1.2.2 Relacionándose socialmente en *Internet*: *Social network sites* o sitios de interacción social

Desde la construcción teórica de la denominada Web 2.0 [O'R04] se plantea un escenario digital centrado en la interacción interpersonal con importantes efectos sociales. Son siete los principios de las aplicaciones Web 2.0 [O'R07]: (i) la Web como plataforma, (ii) el aprovechamiento de la inteligencia colectiva, (iii) los datos son los siguientes elementos a integrar físicamente en los dispositivos, (iv) el fin de ciclo de las actualizaciones, (v) el modelo programación ligera[1], (vi) el software no limitado y (vii) las experiencias enriquecidas del usuario. En particular, tienen un especial interés en la dimensión política de las redes sociales, la web como plataforma, la inteligencia colectiva, los datos como principal activo, el software descentralizado en más de un solo dispositivo y finalmente las experiencias enriquecidas de usuario.

A partir de ese momento, los usuarios pueden intercambiar contenidos libremente, por lo que la generación y distribución de éstos se hace más transparente y en distintos sitios web. La diversidad de contenido y formatos (documentos, texto, imagen, audio, vídeo) crece a medida que se crean nuevos servicios que permiten crearlos y publicarlos de manera sencilla, o hacerlos accesibles desde diferentes dispositivos, desde cualquier lugar y en cualquier momento. Este proceso se acentúa con la proliferación de teléfonos inteligentes o *smartphones*. Los usuarios se convierten en sujetos activos, creando, opinando, participando, relacionándose y compartiendo información, haciendo de este escenario un nuevo entorno dinámico, participativo y colaborativo.

De entre todos los tipos de servicio, los más relevantes son: (i) RSS - Suscripción Realmente Simple o *Really Simple Syndication*), (ii) *wikis* o sitios para generación de contenidos de manera colaborativa [Cha07b][2]), (iii) *blogs* para la publicación de artículos más o menos breves de manera periódica[3] , (iv) sitios web para la gestión y publicación de fotos[4], (v) sitios web para publicación de vídeos[5], (vi) *folcsonomía*[6] que permite etiquetar y clasificar los contenidos que cada uno produce y/o consume [Gru07], y finalmente (iv) redes sociales.

[1]Paradigma que busca sustituir los complejos diseños por otros más pragmáticos basados en la simplicidad y fiabilidad de aplicaciones web.

[2]Ejemplo de *wiki*: Wikipedia (`http://es.wikipedia.org`)

[3]Ejemplo: `http://www.chap:introduccionwordpress.com`

[4]Ejemplos: *Flickr* (`http://www.flickr.com`), *500px* (`http://www.500px.com`)

[5]Ejemplo: *YouTube* (`http://www.youtube.com`)

[6]Folcsonomía es cualquier indexación social por medio de etiquetas simples en un espacio de nombres en el que no existe relación jarárquica o de otro tipo predeterminada entre dichas etiquetas.

La red social *MySpace* nace en 2003, Facebook en febrero de 2004, mientras que *YouTube* lo hace en 2005. En tan solo una década el despegue de las denominadas *"redes sociales"* en *Internet* ha supuesto una importante novedad en el uso de las nuevas tecnologías, tanto por el modo en que se utilizan, como la utilidad y las consecuencias que acarrean en el comportamiento de las personas.

Las redes sociales o sitios de interacción social de la web 2.0 permiten construir perfiles públicos o semipúblicos a los usuarios para poder conectarse a otros perfiles e intercambiar o compartir con ellos distintos tipos de información, con un carácter transaccional. La naturaleza y nomenclatura de estas conexiones o enlaces entre nodos pueden variar de un sitio a otro [BE07].

Sin embargo, la acepción del término *"red social"* desde el campo de la tecnología en cierto modo oculta que en realidad dicho concepto es anterior a las redes sociales digitales o redes sociales *online*. Desde la Sociología y la Ciencia Política, han tratado de dar algunas respuestas e interpretaciones sobre esta materia en su dimensión social, relacional o política, por citar algunas. Sin embargo, ha sido *Internet* el canal por el cual las nuevas aplicaciones web denominadas a sí mismas como "redes sociales"han permitido que se establezcan nuevos *vínculos digitales* entre *personas o identidades digitales*. Sin embargo, la dimensión social va mucho más allá de la mera conectividad electrónica, tal y como se ha descrito anteriormente.

Los sitios de interacción social, o *social network sites*, son aplicaciones o servicios basados en web que tienen como propósito habilitar una conectividad social desde un enfoque personalizado. Aunque muchos de estos sitios web de interacción social están concebidos para ser abiertos, otros están dirigidos a grupos más especializados, según distintos criterios: geográficos, profesionales, religiosos, según intereses o aficiones personales. Estas plataformas ofrecen distintas prestaciones para personas, grupos y organizaciones, que se enumeran a continuación [BE07, KK09]:

- Dotar de un perfil público o semi-público en un sistema acotado definido bien por la propia plataforma o bien por el propio usuario a través de políticas de privacidad.

- Registrar información personal del propio usuario que permita la incorporación de personas con las que tenga en común algunas de sus características.

- Construir una lista de usuarios con los que compartir al menos una conexión, a los que es posible acceder para realizar transacciones con ellos.

- Visualizar y atravesar su lista de conexiones y acceder a aquellas listas creadas por otros usuarios dentro del sistema.
- Contar con mecanismos de comunicación interpersonal, ya sea mediante mensajes privados instantáneos, o bien mensajes colectivos o de forma pública, a través de un *Muro* o herramienta similar. Un *Muro* actúa como un tablón en el que es posible publicar información, ya sea una opinión, un comentario, una pregunta o una nota de actualidad. La información puede publicarse a título individual o en representación de una organización o institución.
- Compartir material multimedia: imagen, audio, vídeo, así como ofrecer enlaces a otros contenidos ya sean propios o de terceros.
- Elaborar y acceder a otros servicios basados en aplicación, que se pueden construir a partir de tecnologías abiertas o bien propias de la plataforma (mediante API[1]).

A través de este tipo de herramientas o plataformas, una persona puede tratar de acceder a otra u otras que forman parte de una red. Cuando se trata de conectar con personas desconocidas, esta actividad se denomina *networking*, con lo que el proceso de inicio de la relación adquiere un matiz especial [BE08].

A diferencia de los foros de discusión, los sitios de interacción social o redes sociales presentan algunas diferencias notables, que pueden tener implicaciones sobre la dinámica en dicha plataforma:

1. En primer lugar, son un entorno con moderación muy baja o *sin moderación*, esto es, no existe la figura de un moderador que realiza un control explícito acerca de la idoneidad o cualquier otro criterio sobre lo que se publica, restringiendo a priori o a posteriori comentarios de los participantes. Los sistemas auto-moderados han sido analizados para estudiar la calidad de los comentarios a partir de sistemas de puntuaciones que otorgaban los participantes [LR04, DlP14]. También ha sido objeto de estudio la tesis de que estos foros podrían representar una esfera pública o un *público virtual* [Hal01, Poo05, Bao00], con las matizaciones específicas sobre lo que representa una web comercial frente a un espacio público.
2. No existen comentarios anónimos como tales, sino que son personales de cada individuo, más allá de que un individuo a través de su *identidad digital* se corresponda con una persona física real o creada artificialmente.

[1] API: *Application Programming Interface* o Interfaz de Programación de Aplicaciones

En el ámbito de la política, durante la campaña electoral norteamericana a las elecciones presidenciales de 2008, la organización del candidato Barack Obama realizó un uso intensivo de *Internet* en su relación con los ciudadanos, y en particular a través de su perfil en *Facebook*, lo que provocó la mayor captación de fondos de la historia en Estados Unidos a través del mayor número de personas. Años antes, el candidato Howard Dean [Hin05] había sentado los precedentes, revolucionando el proceso de financiación de los candidatos y por tanto la dinámica electoral norteamericana, aunque sin alcanzar el grado de movilización de personas de la campaña de Obama. Desde entonces, distintas iniciativas del mismo tipo han tenido lugar en otros procesos electorales, en especial durante las elecciones en Reino Unido de 2010. En España, este tipo de estudios aunque se han planteado, aún no gozan de un desarrollo politológico-tecnológico extendido, lo que propicia la aparición de nuevos trabajos en este ámbito.

La investigación que se plantea tiene un fuerte carácter exploratorio y por tanto descriptivo de un fenómeno, el político-digital, que parece albergar cierta relación con la realidad del comportamiento político en España, que en los últimos tiempos contiene rasgos de una creciente desafección y polarización de los posicionamientos hacia la política y los partidos [VB11, VJ12, RVLET12].

1.2.3 **Superando la dicotomía real-virtual: *social informatics***

Durante bastante tiempo se ha planteado una dicotomía entre el contacto físico de una persona y el contacto a través de dispositivos tecnológicos. Resulta difícil discernir entre qué es *físico* y qué es *virtual* cuando la experiencia física ya incluye una doble faceta material y *online*. Los efectos mutuos entre lo tecnológico y lo físico han quedado de manifiesto en el caso de la primavera árabe, el 15M o el movimiento *Occupy Wall Street*, en los que la tecnología ha ocupado un lugar predominante como instrumento al servicio de la movilización social.

Durante los últimos años la relación entre *Internet* y la sociedad se ha abordado desde distintas perspectivas, como la investigación sobre *Internet*, las TICs[1] y la sociedad, los medios sociales, la teoría de la sociedad de la información y otros, entre ellos la denominada informática social (del inglés *social informatics*) [KRS05]. Este término, acuñado por Kling *et al*. se define como *el estudio interdisciplinar sobre el diseño, usos y consecuencias de las TICs que tienen que ver con los contextos institucionales y culturales*. Aborda la relación entre las TICs y la sociedad como un proceso dialéctico de constante interacción mutua. Por un

[1] Las TICs, o *ICTs* en su versión anglosajona, es el conjunto de tecnologías de la información y la comunicación.

lado, como un proceso en el que los seres humanos diseñan y construyen tecnologías que están influidas por las necesidades o intereses de la sociedad; y por otro lado, un proceso en el que esas nuevas tecnologías dan forma la sociedad, transformándola.

Esta informática social tiene un marcado carácter empírico, por encima del teórico; corpus que va construyendo principalmente a través de las contribuciones experimentales.

Mantienen un planteamiento crítico, esto es, no se acepta sin cuestionar los propósitos y creencias que se dan por descontadas por las organizaciones y personas que ponen en marcha el desarrollo de nuevas tecnologías. En consecuencia, esta postura crítica, no asume el determinismo tecnológico.

Desde el punto de vista metodológico, esta *social informatics* cuestiona la idoneidad de los métodos clásicos de análisis social. Distintas posturas confrontan sus puntos de vista, pero en términos generales se asume que nuevos métodos de investigación que han nacido en la era de *Internet* pueden y deben complementar a los métodos tradicionales [BRB02, CHJ03, Fuc07]

Se trata de un campo amplio, abierto y por definir. La complejidad a la hora de definir sus límites estriba en gran medida en lo cambiante de los fenómenos que describe y de la mirada multidisciplinar que proyecta sobre ellos [Hun05, LS05, Saw05].

1.3 **Motivación**

La sociedad actual es una sociedad hiperconectada, una *sociedad-red* [Cas09] de personas crecientemente contectadas en la que las redes de comunicación actúan como una infraestructura de dispositivos y servicios avanzados para conectar personas, información, opiniones, voluntades e intereses de toda naturaleza, desde las económicas hasta técnicas, sociales o afectivas.

Frente a unos años donde lo político y la política han venido siendo cuestionados y deslegitimados [Lue09], el mundo digital presenta oportunidades para cambiar la calidad democrática de la sociedad y de sus instituciones. Las redes de comunicación [Cas09] que han proliferado en las últimas décadas no son sino una expresión de una sociedad que vive crecientemente en red. Se antoja especialmente relevante conocer cómo se aglutinan demandas, necesidades, oportunidades y capacidades políticas para poder transformar una realidad política en otra distinta.

En política, resultan especialmente relevantes los mecanismos y procesos de representación y en mayor grado los procesos electorales para la conformación

de gobiernos. Estos procesos han venido adquiriendo una notable dimensión comunicativa desde la aparición de los medios de comunicación de masas, y con la aparición de *Internet* esta comunicación se ha transformado radicalmente. El carácter horizontal de la comunicación ha venido incrementando un carácter más democrático y plural. Instituciones políticas como los gobiernos, las administraciones o los partidos políticos han entendido que es posible establecer una interacción política más cercana con los ciudadanos, con un carácter a priori más dialogante, constructivo y participativo.

Por otro lado, las redes sociales *online* han pasado a ser un instrumento para el ocio de universitarios estadounidenses, como es el caso de *Facebook*, a convertirse en una herramienta de uso cotidiano en el ámbito personal y profesional. Muchas empresas, organizaciones e instituciones públicas comenzaron a darles un uso dirigido principalmente con fines de promoción de una imagen actual, moderna y dinámica, y han continuado haciendo un uso comercial a partir de los servicios que estas redes ofrecen. La esfera digital, que nació a través de portales web, floreció con los blogs y se ha desarrollado exponencialmente a través de redes sociales *online*. Estas redes sociales digitales actúan como una tecnología habilitante que en lo político, partidos y otros actores se comunican directamente y de manera bidireccional con los ciudadanos.

La confluencia de estos dos procesos -uno sociopolítico y otro tecnológico- nos plantea el reto de conocer cómo es el debate o *discusión* política (en el sentido anglosajón del término *discuss*) que se establece entre los ciudadanos y los partidos en ese nuevo escenario que constituyen las redes sociales *online*. Cabe preguntarse cómo los partidos políticos llevan a cabo su comunicación en sitios de interacción social *online* y por tanto conocer si la nueva esfera digital es un espacio de deliberación democrática con mayor pluralidad y riqueza que el espacio público ya conocido. O bien si estamos ante una nueva dimensión en la que simplemente se replican las mismas dinámicas comunicacionales anteriores.

En cuanto a la política en España, desde su instauración en la transición, el sistema de partidos ha mantenido un comportamiento bipartidista [Lin04, Lle98], en el que dos partidos principales (*Partido Socialista Obrero Español*, PSOE, y *Partido Popular*, PP) han obtenido sistemáticamente un porcentaje de voto superior al setenta porciento en elecciones generales y autonómicas, y con la excepción de la fase inicial de la democracia se han alternado en el ejercicio del gobierno durante casi un cuarto de siglo. Estos dos grandes partidos conviven con otros dos de ámbito no estatal (*Convergencia i Unió*, CiU, y *Partido Nacionalista Vasco*, EAJ-PNV) con una presencia institucional en sus dos territorios de origen - Cataluña y el País Vasco respectivamente-, con una cada vez menor representación de un partido de ámbito estatal, configurada alrededor

de Izquierda Unida o cualquiera de sus denominaciones a lo largo del tiempo.

Este modelo de partidos se pone en cuestión con la irrupción de varias fuerzas políticas de reciente creación y que cuentan con una reducida representación institucional hasta la fecha [PG13]. A partir de 2014, dos nuevas formaciones, *Podemos* y *Ciudadanos*[1], aparecen en la escena política nacional en unas circunstancias en las que existe una sustancial desafección política hacia los partidos [dIS15, VB11].

1.4 Reto general

A la luz de lo anterior, este escenario plantea un fenómeno que tiene múltiples facetas. No pretende únicamente describir un nuevo fenómeno de movilización de la sociedad civil y los actores políticos enfocándolo desde la perspectiva de la acción política, sino que aborda la cuestión como un fenómeno poliédrico.

De una parte, se plantean varias cuestiones sobre las estrategias de comunicación en la esfera digital, sus efectos y consecuencias. En primer lugar, destacan aquellas relacionadas con la calidad democrática de la interacción digital que sucede en estas nuevas redes de interacción. Esto es, si este fenómeno se *corresponde* con un espíritu de debate cívico basado en el pluralismo y la discusión abierta, o si, por contra, carece de esta guía y en realidad lo que sucede es que es una réplica de las estrategias convencionales de comunicación de los partidos políticos a través de un nuevo canal tecnológico.

En este sentido, la presente investigación tratará de constatar las posibilidades efectivas de comunicación por los partidos en una arquitectura de la intercomunicación [O'R07] como la que ofrece la web 2.0 a través de los sitios de interacción *online* y su capacidad de influencia más allá de la pura transmisión de información, esto es, en los estadios siguientes hacia la movilización individual y colectiva.

En segundo lugar, respecto a la eficacia comunicativa de dicha comunicación, resulta pertinente conocer las estrategias comunicativas en la medida en la que éstas pueden contirbuir a la interpretación y creación individual y colectiva de significados. Esta respuesta puede ofrecer una mirada del fenómeno desde lo político (el sistema político, la democracia), desde lo tecnológico y, por último, desde el nivel estratégico y operativo de los partidos políticos.

En tercer lugar, la información que emerge a través de dispositivos tecnoló-

[1] *Ciudadanos-Partido de la Ciudadanía*, es un partido fundado en 2006 a partir de la plataforma ciudadana *Ciutadans de Catalunya*, inicialmente con ámbito de acción política en Cataluña, y a partir de las elecciones municipales de 2011 en gran parte del resto de España.

gicos supone un terreno fértil sobre el que construir conocimiento a partir de los contenidos que tanto los partidos políticos como los ciudadanos vierten en las nuevas redes sociales *online* que son expresión de sus intereses, necesidades y opiniones. Desde las ciencias de la computación, se plantea como desafío poder abordar el análisis y extracción de conocimiento de manera automatizada de dicha información expresada.

Finalmente, desde la estrategia y operaciones de los partidos políticos, los agentes sociales, los movimientos sociales y los ciudadanos, la presente investigación pretende ofrecer algunas respuestas a los actores políticos, que les permita realizar un diagnóstico y establecer recomendaciones de acción. En esta dirección, permitirá identificar y reconocer los temas estratégicos que se utilizan durante las campañas electorales con el propósito de que la comunicación sea efectiva y eficaz.

A estas cuestiones pretende dar respuesta la presente investigación, motivo por el cual resulta necesario articular una hipótesis fundamental que conducirá este trabajo.

1.5 Hipótesis y objetivos

A partir del reto general planteado, se ha formulado la hipótesis sobre la que se fundamenta esta investigación, que se enuncia a continuación:

> «*Se puede describir el modelo de acción de los principales partidos políticos en España en sitios de interacción social determinando el comportamiento en su acción comunicativa a través de métodos de recuperación de información, minería de datos y técnicas de aprendizaje automático.*»

A partir de esta hipótesis, el objetivo principal que se busca alcanzar en esta investigación es el siguiente:

Objetivo principal 1. *Desarrollar y validar un sistema que describa el comportamiento comunicativo de los partidos políticos en sitios de interacción social online.*

Con el fin de lograr este objetivo central, se plantean distintos objetivos específicos, cumplidos los cuales harán que se logre el objetivo de esta investigación y en consecuencia poder a validar la hipótesis fundamental.

Objetivo específico 1. *Definir un modelo de categorización de la acción comunicativa que ocurre en los sitios de interacción digital de los partidos.*

Objetivo específico 2. *Elaborar un modelo que represente la información de la acción comunicativa en dichos sitios de interacción social.*

Objetivo específico 3. *Diseñar y construir un método que permita validar la categorización de la acción comunicativa generada en los sitios sociales de los partidos políticos.*

Para el cumplimiento del primero de los objetivos específicos será necesario elaborar un modelo que responda a las principales dimensiones de la acción comunicativa de los partidos políticos. A continuación, el segundo objetivo específico responde a la necesidad de diseñar un modelo que represente de manera adecuada la información derivada de esas dimensiones de la acción de los partidos. Para finalizar, el tercer y último objetivo específico consiste en crear un método que, a partir del modelo de representación previamente elaborado permita categorizar la acción comunicativa en los términos definidos en el primer objetivo. Además, se deberá validar el propio método de clasificación.

Para poder alcanzar estos objetivos específicos, ha sido necesario descomponerlos en una serie de objetivos operacionales, que se enumeran seguidamente:

Objetivo Operacional 1. *Diseñar y construir un sistema que permita recoger de manera automática la información que representa la comunicación que tiene lugar en los sitios de interacción digital de los partidos.*

Objetivo Operacional 2. *Establecer un modelo de categorización de mensajes publicados en muros políticos.*

Objetivo Operacional 3. *Generar un modelo normalizado que represente la información de los mensajes.*

Objetivo Operacional 4. *Desarrollar y validar un modelo que permita clasificar mensajes en muros digitales de partidos políticos.*

Objetivo Operacional 5. *Establecer un modelo para categorizar las conversaciones políticas en muros políticos.*

Objetivo Operacional 6. *Generar un modelo normalizado que represente la información de las conversaciones.*

Objetivo Operacional 7. *Desarrollar y validar un modelo que permite clasificar conversaciones generadas en los muros digitales de partidos políticos.*

Objetivo Operacional 8. *Definir un método que permita identificar las interdependencias entre las distintas variables que caracterizan a la acción comunicativa.*

Estos objetivos operacionales son las pautas a seguir durante la investigación. El cumplimiento de todos ellos, supondrá el logro del objetivo principal de la investigación y por tanto, la validación de la hipótesis fundamental anteriormente expuesta.

1.6 Metodología de la investigación

La aplicación del método científico en esta investigación supone que la metodología a utilizar responda a dos aspectos claves relacionados con el proceso de experimentación y con la hipótesis, a saber:

- La **reproducibilidad** de la investigación nos permitirá poder alcanzar resultados similares a partir de la misma información de partida, sobre cualquier experimento de la propia investigación en otro momento. Esto será posible en la medida en que los resultados son compartidos dentro de la comunidad científica.

- La **falsabilidad** o **refutabilidad** es la facultad de que una afirmación científica, a través de un contraejemplo, pueda ser rebatida mediante experimentación u observación. Dicho de otro modo, es posible refutar la hipótesis validada en una investigación mediante la realización de experimentaciones que pudieran ofrecer resultados diferentes a los alcanzados por la investigación inicial.

El método de investigación tiene carácter experimental y la técnica de investigación social a utilizar es una variante de la observación participante. Este tipo de técnicas se caracterizan por que el investigador se integra en el contexto

del objeto investigando, formando parte de su entorno, experiencia y realidad [TB98]. Por tanto, las fuentes de información sobre las que se trabajará serán de carácter primario. En el caso presente, la inmersión se produce en el ámbito digital de *Internet* y las redes sociales *online*.

Seguidamente se detallan las tareas a realizar, y que constituyen el centro de la metodología de investigación:

1. **Identificación y descripción del problema concreto**. Para abordar esta tarea, será necesario aproximarnos al fenómeno a estudiar con el fin de identificar el objeto de la investigación. Esta tarea incluye su definición exacta en los términos que la investigación precisa, dejando al margen aquellos que resulten accesorios o menos relevantes. La definición del problema conlleva igualmente plantear un marco teórico sobre el que adquirir el conocimiento básico para investigar dicho fenómeno.

   Por todo ello, resulta necesario conocer y comprender las principales aproximaciones teóricas y prácticas al fenómeno y al objeto de la investigación, a fin de abordarlo de manera adecuada y así poder construir o ampliar el conocimiento científico. Principalmente, se hará uso de fuentes secundarias de información, procediendo de las más genéricas al fenómeno hacia las concretas que nos permitan aproximarnos a sus detalles.

2. **Diseño de la investigación**. Durante esta etapa se seleccionan las técnicas más adecuadas, definiendo el universo muestral y las distintas muestras. Esta actividad incluye, además, la propuesta de diversos experimentos a partir de los cuales se obtendrán los datos necesarios que posteriormente se analizarán. En particular, nuestra investigación incluye el planteamiento de una solución basada en *software* que permitirá la adquisición de dichos datos.

3. **Recolección de información**. Mediante la solución propuesta se recogen los datos en los términos establecidos y se analiza la calidad de dicha información en relación al problema a investigar.

4. **Análisis de los datos y obtención de resultados**. A partir de los datos obtenidos, se aplican distintas técnicas de análisis, en nuestro caso de carácter cuantitativo. Pese a que gran parte de la información que se procesa es cualitativa, las técnicas aplicadas son de análisis cuantitativo, por lo que procederá un proceso que transfiera entre dichos tipos de información de manera precisa.

5. **Interpretación de los resultados**. Los resultados obtenidos deberán ser contextualizados en relación con el fenómeno que se está investigando, de manera que sea significativo con respecto al conocimiento previo en ese área. Dado que el conjunto de experimentos realizados es amplio, será necesario agrupar y comparar los resultados para poder realizar interpretaciones adecuadas.

6. **Diseminación de los resultados**. Esta tarea incluye la compartición y contraste de los resultados alcanzados en la investigación con el propósito de que la comunidad científica tenga constancia de las aportaciones realizadas a partir del nuevo conocimiento elaborado.

1.6.1 Una investigación experimental

Como se ha anticipado previamente, una de las características de la presente investigación es su carácter experimental, frente a los métodos clásicos en investigación electoral basados en otras técnicas. La principal ventaja de los experimentos sobre la técnica de los cuestionarios es la capacidad del investigador para aislar y testear los efectos de componentes específicos de los mensajes políticos [Iye10]. Así, se evita que a través de un cuestionario se expresen reflexiones, opiniones e información distinta a la que una persona a través de su perfil digital manifiesta de manera directa y sin intermediaciones. Los cuestionarios adolecen de los denominados *errores de memoria* que incluyen los indicadores basado en el recuerdo sobre la exposición a determinadas historias de las noticias [Gun87] o las fuentes de las noticias [PZ93].

Las campañas engloban la concatenación de mensajes, canales y fuentes, pudiendo todos ellos influir en la audiencia. Mediante experimentación es posible aislar las explicaciones, sean de manera verbal o en forma de indicaciones audiovisuales.

Otro problema de los cuestionarios es que la exposición a medios autoreportada es típicamente endógena. Así, está comúnmente aceptado que quienes dicen leer periódicos o ver las noticias en televisión difieren sistemáticamente de los que manifiestan que no. Esto se ha acentuado con la expansión de *Internet*. El problema del carácter endógeno tiene muchas derivaciones en la investigación de la comunicación política. Resulta muy frecuente que la exposición reflejada por una persona depende de las predisposiciones políticas, frente a la exposición mediática usual.

Aunque la experimentación directa reduce riesgos en la obtención de información, el uso de tecnología digital en investigaciones experimentales presenta

algunas objeciones a priori. Mientras que, por un lado, el investigador dispone de herramientas para administrar y controlar las manipulaciones, a un público también más controlable, por otro lado se enfrenta a un entorno mediático mucho más alterado en el que la exposición a contenidos políticos se produce por casualidad y no de manera voluntaria [BI08]. Como resultado, el tratamiento experimental dependerá inevitablemente de las preferencias políticas, y estimar los efectos de dicho tratamiento implicará el uso de herramientas estadísticas mucho más potentes. Dado que esta investigación no describe las consecuencias en cuanto al comportamiento al voto, no se ve afectada por estas implicaciones.

1.6.2 Consideraciones de seguridad y privacidad

A lo largo de la investigación se accede a información de personas a través de perfiles digitales en la red social seleccionada, *Facebook*. A la hora de realizar este proceso de captura de información, se exige que las aplicaciones informáticas que se han construido *ad hoc* y el soporte de almacenamiento de los datos gestionen esta información de manera absolutamente confidencial. Por todo ello, la información recogida ha sido registrada de manera impersonal y por lo tanto no vinculada nominalmente con ninguna persona.

A los solos efectos de representación en este documento, de los comentarios y opiniones vertidas en perfiles digitales de los partidos políticos, se estima adecuado que dichas opiniones se asocien nominalmente a sus autores en la medida en que los contenidos publicados en los muros de tales partidos políticos se consideran información pública y por tanto accesible a cualquier persona.

1.7 Diseño general

En esta sección se presenta el diseño general de la investigación que conducirá al cumplimiento de los objetivos y, en consecuencia, la prueba de la hipótesis anteriormente establecida. Como muestra la Figura 1.1, el diseño consta de una serie de componentes, explicados a continuación:

- **Modelo de comunicación política**. El modelo de comunicación abarca las principales dimensiones de la acción comunicativa que desarrollan los partidos políticos.
- **Modelo de representación de la comunicación política**. Dos modelos de representación: (i) de las conversaciones políticas, y (ii) de los mensajes que tienen lugar dentro de las conversaciones políticas.

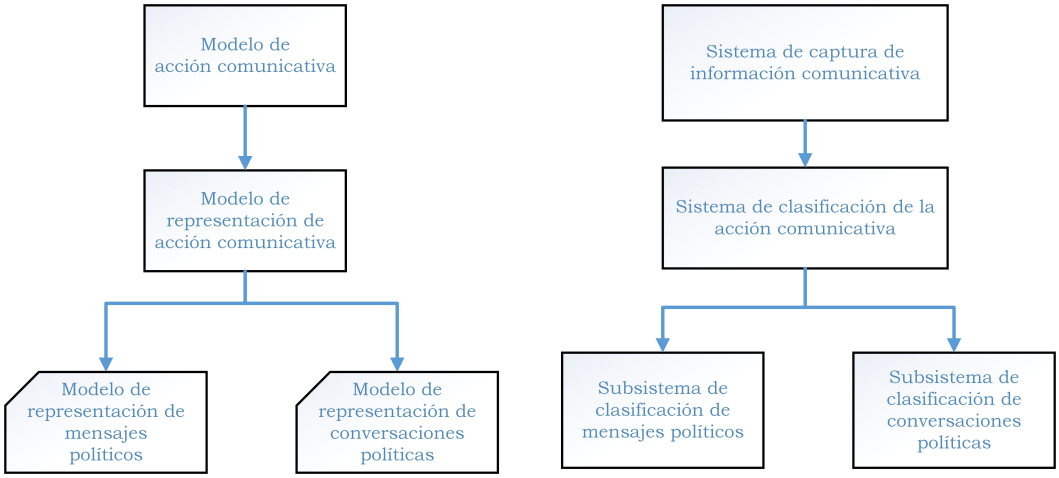


Figura 1.1: Diseño de la investigación.

- **Sistema de captura de información comunicativa**. Este sistema provee de la información publicada en los sitios de interacción social *online* en los términos en que el modelo de representación de la comunicación política tiene definidos.

- **Sistema de clasificación de la comunicación política**. Este módulo incluirá dos subsistemas que permitan clasificar, respectivamente, las conversaciones políticas como los mensajes vertidos en dichas conversaciones en base a los modelos de comunicación política establecidos para cada uno de ellos.

    1. **Subsistema de clasificación de mensajes políticos**. En primer lugar, será capaz de identificar los niveles de las funciones informativa y movilizadora. Del mismo modo, reconocerá el grado de debate político que aporta el mensaje, así como el perfil de referencialidad de los contenidos. También será capaz de identificar automáticamente el nivel de personalización, junto al nivel de meta-comunicación. Con respecto a la orientación estratégica, el módulo permitirá reconocer los marcos primario y secundario de la comunicación, así como la eficacia del marco de responsabilidad si fuere el caso, además del nivel de campaña negativa de un mensaje determinado.

        Para ello, utiliza un modelo de aprendizaje automático, entrenado previamente con un conjunto de datos formado por mensajes iniciales de las conversaciones, que categorizará cada uno de los *posts* en relación a las clasificaciones citadas.

2. **Subsistema de clasificación de conversaciones políticas** El subsistema de clasificación de las conversaciones o discusiones políticas se establece a dos dimensiones de la acción comunicativa: el nivel funcional y el nivel de estrategia comunicativa. En la primera dimensión se incluyen la función informativa, la función de movilización, el nivel de personalización y el nivel de meta-comunicación. En la segunda se consideran los marcos primario y secundario, la eventual eficacia del marco de responsabilidad y finalmente el nivel de campaña negativa de la conversación.

   Con este fin emplea un modelo de aprendizaje automático, que previamente ha de ser entrenado utilizando un conjunto de datos compuesto por las conversaciones políticas, y que permitirá clasificar cada una de las conversaciones en relación a cada una de las distintas clasificaciones enumeradas anteriormente.

1.8 Estructura de la tesis

Hasta este punto se ha presentado el tema central de la tesis, formulada la hipótesis y descritos los objetivos. Posteriormente se ha propuesto una arquitectura general que de respuesta a tales objetivos y el diseño para llevar a cabo la investigación. El resto del documento queda organizado en los siguientes capítulos:

Capítulo 2. Revisión de la literatura. A lo largo de este capítulo se presenta el estado de la cuestión en las materias de la investigación. En primer lugar, los enfoques sobre la nueva esfera democrática digital. Seguidamente se muestran las estrategias, tanto las orientadas a la competición electoral como las estrategias de comunicación durante estos procesos democráticos. Tras este bloque, se exponen los métodos de procesamiento de datos y texto, filtrado de contenidos.

Capítulo 3. Modelado y procesado inicial de los datos. Esta sección de la tesis introduce el sitio de interacción social sobre el que se desarrolla la investigación, *Facebook*. A continuación, la sección detalla el proceso de captura y procesamiento de la información necesaria, así como la construcción de los conjuntos de datos sobre los que se realizará la experimentación. Continúa el capítulo describiendo el proceso de etiquetado de los conjuntos de datos, la caracterización de dichos conjuntos de datos a través de atributos que serán de fundamental utilidad en las etapas posteriores.

Capítulo 4. Categorización automática de mensajes y conversaciones políticas. Una vez descritos en el capítulo anterior los conjuntos de datos y extraídas sus características definitorias será necesario desarrollar una catego-

rización automática de los mensajes y las conversaciones políticas. Para ello, este capítulo comienza describiendo los algoritmos de aprendizaje supervisado utilizados durante la experimentación, antes de pasar a la validación empírica que permite evaluar los resultados en ambos conjuntos de datos.

Capítulo 5. Mejora del modelo. A lo largo de este capítulo se realiza un análisis entre las distintas dimensiones comunicacionales además de otras técnicas de procesamiento de información. Sobre estos dos pilares se proponen mejoras sobre el modelo de categorización obtenido en el capítulo anterior, de modo que permite clasificar las interacciones políticas con mayo eficacia.

Capítulo 6. Conclusiones. La tesis finaliza presentando las principales contribuciones realizadas repasando de nuevo la hipótesis fundamental y los correspondientes objetivos. A su vez, también plantea las limitaciones del presente trabajo, junto a futuras direcciones sobre las que continuar a partir de la presente investigación, y unas consideraciones finales al trabajo presentado.

«La identidad se encuentra no solo en la raiz, sino también en la relación»

Édouard Glissant, *Poétique de la relation* (1928-2011)

CAPÍTULO

2

Revisión de la literatura

La presente investigación ofrece una mirada múltiple a un fenómeno que a su vez cruza varias disciplinas: política, comunicación e *Internet*. Por ello, tras haber definido el marco general, en este capítulo se presenta un estado de la cuestión de los principales avances científicos y la praxis asociada a dichos campos. Sin ánimo de ser exhaustivos, se señalan aquellos aspectos de las materias referidas que están directamente relacionados con la propia investigación.

En primer lugar, la comunicación ha vivido en las últimas décadas una era dorada en la que ha venido alcanzando cotas cada vez mayores de protagonismo en el ámbito del ejercicio del poder y de la política, con especial notoriedad la comunicación de los partidos políticos e instituciones públicas. Todos ellos han desplegado diferentes estrategias y técnicas al servicio de sus resultados partidistas, institucionales o gubernamentales.

En paralelo a esta evolución, el desarrollo de tecnologías de información y comunicación ha tenido un enorme impacto en la sociedad, y el espacio de lo político, de lo público, se ha visto afectado por ello.

Estas tecnologías, al hilo del desarrollo de *Internet* como columna vertebral sobre la que enormes cantidades de información fluyen instantáneamente, han propiciado nuevos espacios de expresión política. Se dispone de información cada vez más variada, rica, sofisticada que manifiesta la voluntad de libertad de expresión y por tanto supone un verdadero escenario en el que tiene lugar un nuevo estilo de hacer política. Estas nuevas formas de transmisión de infor-

mación y conocimiento han encontrado cauce a través de canales novedosos para la comunicación en los tiempos recientes, y especialmente a través de las denominadas redes sociales *online*.

Por todo ello, desde las ciencias de la computación, la gestión de dicha información supone un auténtico desafío en la medida en que la cantidad de información ha crecido de manera exponencial, con una diversidad de fuentes en gran medida heterogéneas. Al mismo tiempo que aumenta la velocidad a la que crecen los datos, se ha de hacer frente a la heterogeneidad de los tipos y formatos de información. Adicionalmente a este crecimiento de la información, se puede ofrecer respuesta al reto que supone comprender de manera más o menos automatizada, al significado, propósito o estrategia que subyace a este continuo flujo de información.

Para tal fin, el capítulo se ha organizado en torno a estas principales materias. En primer lugar, la sección 2.1 describe los cambios que *Internet* ha producido en un ámbito particular, el de la política. Además, se analizan experiencias de la denominada *ciberpolítica*, en confluencia de ciudadanos, actores políticos a través de redes. Tras trazar un recorrido de los partidos políticos y el cambio de la información política en el nuevo escenario, la sección en su parte final aborda la cuestión de la democracia en una nueva esfera digital.

Seguidamente, la sección 2.2 trata los distintos paradigmas de los estudios electorales cuyo propósito es conocer el comportamiento de los ciudadanos a al ahora de votar. Tras ello, repasamos los principales rasgos de las campañas electorales.

A continuación, la comunicación política y electoral se describe en la sección 2.3. Se realiza un recorrido por la transformación de los modelos de comunicación de los partidos, y en particular se describe el uso de la red social *Facebook* como canal de comunicación política. La sección termina planteando el análisis de marcos como esquema para la interpretación de la comunicación política. Durante estas primeras secciones se presenta un estado de la cuestión sobre el fenómeno de la comunicación política en un entorno, el digital, que pone sobre la mesa transformaciones que tienen un potencial impacto.

A continuación, la sección 2.4 revisa a los métodos que son utilizados generalmente para el procesado automático de datos y textos. Por su parte, la sección 2.6 desarrolla algunas de las técnicas de filtrado de contenidos y textos. Para finalizar el capítulo, la sección 2.7 sintetiza los conceptos que aparecen a lo largo de este capítulo.

2.1 Internet y democracia 2.0: entre la emancipación democrática y la replicación de espacios

El devenir de Internet como una plataforma tecnológica con capacidad para comunicar y compartir información ha permitido el crecimiento de determinados negocios, el nacimiento de una nueva economía y también ha tenido consecuencias sobre los individuos entre sí y dentro de la sociedad. Sin embargo, en relación a los procesos sociales y políticos, existen determinadas limitaciones que cuestionan su neutralidad.

En este terreno, los partidos se han ido incorporando a las dinámicas de Internet creando sitios web en los que informan de sus programas, dan a conocer sus estructuras orgánicas, sus propuestas políticas o información sobre sus candidatos. Posteriormente han añadido a estas secciones, otras prestaciones más orientadas hacia la movilización del electorado o, en el caso de los partidos en Europa, de sus simpatizantes y afiliados. Entre estas nuevas funciones cabe señalar el envío masivo de correos electrónicos, los servicios de notificación por SMS y en las últimas elecciones grupos de *Whatsapp*[1] [VBVAdV10, AC08, Smi11].

Los trabajos sobre las campañas *online* en distintos países muestran que se ha producido un cambio hacia un modo más directo y estrecho de comunicación de los partidos con los votantes que puede, además, fomentar oportunidades de participación para hacer que los ciudadanos se involucren más en los procesos de deliberación democrática y de manera especial en la elección de sus representantes políticos. Existe un cuerpo de experiencias empíricas durante la primera década del siglo XXI, pero no queda totalmente evidenciado el peso de *Internet* sobre las campañas electorales [BCR06, AC08, ACC08].

Más allá de Estados Unidos, las experiencias previas a la web 2.0 son expresiones muy limitadas, con un enfoque mucho más cauto y conservador por los partidos y candidatos durante las campañas electorales. Así, la acción electoral *online* se ha venido centrando en reproducir la información ya elaborada para su uso en los medios tradicionales *offline*, para conseguir únicamente una potencial audiencia mayor [Col01, JPW01, Jac07a]. En parte, estas reticencias se deben al cambio que supone pasar de la era de la televisión, con una gestión y control del mensaje estrictos a un nuevo mundo, el de *Internet*, en el que la regulación sobre la información es mucho más débil y en permanente cambio [GW12].

En cambio, aportaciones recientes muestran la aparición de estrategias de campaña *bottom-up* o *de abajo hacia arriba* [Gue08, TWPL06], especialmente

[1] `http://www.whatsapp.com`

en dos campañas demócratas, las de Howard Dean en 2004 y Barack Obama en 2008, en las que fueron protagonistas la captación de fondos, los *blogs* y el envío masivo de mensajes en forma de *SMS* para apoyar a los candidatos [Gre09, HP09, Nor08]. Para Joe Trippi, director de la campaña digital de Howard Dean, y pionero en este tipo de estrategias, esta transformación consiste en pasar de un modelo basado en la división *nosotros-ellos* a un modelo de base o *grassroots* [Tri05]. Este modelo no es extensible a todos los países en la medida en que fuera de Estados Unidos, tanto los modelos de campaña como las propias estructuras de los partidos están más centrados en el partido que en el candidato.

2.1.1 Ciberpolítica, redes sociales y nuevas movilizaciones

La web 2.0 como plataforma ofrece herramientas discretas que reducen las barreras de acceso, tanto a la exposición a información política, como a la expresión de la opinión propia y al debate político. Estas barreras tecnológicas que existían inicialmente a la publicación de contenidos, se han ido reduciendo con la incorporación de tecnologías como los blogs, *RSS*[1] o la publicación de vídeos en redes sociales como *YouTube* y otros servicios similares. La edición y gestión de contenidos, su publicación y compartición resulta sencilla, en gran parte de sus formatos: texto, documentos, sonido, imágenes, vídeo.

La clasificación que Dahlgren realiza nos permite enumerar hasta 5 formas distintas de interacción política en Internet [Dah05a], a saber:

1. *Portales de gobierno electrónico o e-government*, en los que los gobiernos o sus Administraciones (administración electrónica o *e-administration*) ofrecen una ventana digital en aras a incrementar la transparencia y visibilidad de sus políticas públicas y la eficacia de las mismas, además de promover la participación electrónica, con desiguales resultados.

2. *Dominios de activistas* de causas a partir de las que se generan discusiones auspiciadas por colectivos u organizaciones, generalmente no gubernamentales. Generalmente albergados bajo una infraestructura de sitio web.

3. *Foros cívicos*, donde se produce intercambio de opinión, puntos de vista, generalmente debatiendo sobre aspectos de interés público, por lo que tienen un marcado carácter de orientación hacia la participación.

[1] *Really Simple Syndication* es un formato para sindicar o compartir contenido en la web.

4. *Portales "parapolíticos"*. Este tipo de sitios web propiamente no están diseñados para tratar específicamente asuntos políticos. Sin embargo, los partidos y otros actores políticos pueden hacer uso de estas plataformas, haciendo que sí se incluyan estas cuestiones políticas. Ejemplos de estas plataformas son la redes sociales *Facebook* o *Twitter*[1].

5. *Dominios de medios de comunicación adaptados al espacio digital*, además de los que tienen carácter asociativo o comunitario.

Uno de los elementos de la web2.0 son los sitios sociales de trabajo en red o *social network sites*. Estas redes interpersonales no suponen un fenómeno nuevo [BI08], al menos en investigaciones sobre movimientos sociales [DPD09, Tar05], pero los efectos de estas tecnologías nuevas suscitan un creciente interés científico. Los estudios comparativos a nivel internacional también han puesto la mirada en hasta qué punto las tecnologías de la información y la comunicación podrían aumentar el nivel de competitividad de los partidos menores y sus candidatos para convertir las elecciones en procesos más abiertos, menos predecibles y/o volátiles. Esa esta la tesis de Corrado y Fireston, Morris y Selnow [CF96b, Mor11, Sel98], para quienes Internet podría erosionar el dominio de los actores del *mainstream* [CF96b, Mor11]. Sin embargo, gran parte de la literatura académica plantea un refuerzo efectivo de las tesis normalizadoras según las cuales los grandes partidos y sus candidatos cuentan cada vez más con mejores profesionales que ayudan a poner en marcha campañas mucho más sofisticadas [BD03, GNW03, MR00].

El debate que se establece sobre las nuevas tecnologías abarca tres dimensiones:

1. **El impacto en la opinión pública**. La construcción de la opinión pública es un proceso fundamental, en tanto en cuanto su control y seguimiento permite generar ciertos procesos de enmarcado o *framing* de la realidad. En estos procesos de enmarcado lo primero son las ideas, que son transmitidas por el lenguaje, que presenta y evoca a dichas ideas.

   En la medida en que la opinión pública está conformada como el agregado de juicios individuales, se expresa a través de los medios de comunicación social, es identificada y reconocida mediante sondeos, y finalmente se manifiesta democráticamente por medio de las urnas, al mismo tiempo que es sometida al control de las elites político-institucionales. Por otro lado, la opinión pública discursiva es un proceso en el que un público como colectivo de opiniones individuales que deliberan, se influyen mututamente

[1] `http://www.twitter.com`

y condicionan durante el acto comunicacional [Sam00]. Este acto comunicacional interactivo se enmarca en la sociedad civil [Blu51].

2. **La esfera pública en el ciberespacio**. La esfera pública es "un espacio de discurso, institucional o geográfico, en el que el individuo ejerce como ciudadano, accediendo metafóricamente al diálogo sobre las cuestiones de la comunidad" [Dah05b, Sam00]. Existen varias esferas públicas: una central y varias periféricas. El dinamismo social de las esferas periféricas se ve afectado en tres facetas: la estructural (relacionada con aspectos de regulación y control), la representativa (a través del pluralismo, las tendencias ideológicas y la agenda política) y, por último, la interactiva, esto es, el propio proceso de deliberación.

   Ante esta tercera dimensión surge el debate sobre si *Internet* es en efecto un nuevo espacio de deliberación, ofreciéndose dos conjuntos de visiones contrapuestas. Una perspectiva *ciberoptimista* según la cual la red de redes es una verdadera estructura de oportunidad para la participación y la transparencia institucional, y otra postura *ciberescéptica*, que destaca los aspectos contextuales sobre el potencial tecnológico. Para los escépticos, las condiciones estructurales, socioeconómicas, políticas (cultura política, el sistema electoral, el sistema de partidos, los estilos de campaña, movilización social) condicionan los usos de *Internet*, y por tanto plantean que *Internet* es más un nuevo canal de comunicación multidireccional en el que principalmente se reproducen las mismas dinámicas que fuera de dicha esfera digital.

   En la Tabla 2.1 se sintetizan las principales posturas teóricas:

3. **El carácter de las nuevas movilizaciones**. *Internet* supone un cambio en la propia naturaleza de las movilizaciones. Durante los primeros años del siglo XXI se están produciendo nuevas formas de movilización más descentralizadas, no sujetos estrictamente a los dictámenes de los partidos, sindicatos u organizaciones convocantes [BI08].

   Así mismo, se manifiesta un creciente grado de espontaneidad, con una componente cada vez más personal, principalmente en contextos de crisis, lo que Sampedro denomina un nuevo tipo de "protesta posmoderna", originada en estilos de vida común, y con movilizaciones más de carácter expresivo que sustantivo, tratando de lograr, más que el cambio social, la visibilización de una cuestión [Sam05].

   El surgimiento de una nueva política, más ligada a los valores, de carácter postmaterialista supone un cambio sustancial frente la política de

Tabla 2.1: Cuadro síntesis de las visiones optimistas-pesimistas sobre Internet. Fuente: [L01].

Optimistas	**Pesimistas**
Simone [Sim08] Internet provee medios que favorecen la democracia deliberativa, al permitir, por ejemplo, que públicos alternativos encuentren espacios compartidos.	*Havick* [Hav00]; *Mayer* [May01] La fragmentación de la información dentro de Internet obstaculiza la deliberación.
Dahlgren [Dah05b] Internet extiende y pluraliza la esfera pública en un buen número de vías. Facilita una notable heterogeneidad comunicativa, aunque esto mismo puede generar procesos de fragmentación.	*O´Donnel* [O'D01]; *Steiner* [Ste05] Dentro de la red, se dan espacios que siempre excluyen a algunos grupos (especialmente, a aquellos contrarios a los valores dominantes).
Bennet [Ben03] La red digital ofrece una variedad de formas organizativas, con modelos horizontales y fluidez comunicativa entre los sujetos.	*Galston* [Gal03] Se corre el riesgo de que Internet termine por ser un conjunto de islas de comunicación política en donde, incluso, se formen "ciberguetos".
Friedland, Hove y Rojas [FHR06] Las formas de comunicación en red proveen la forma de conexión entre redes sociales diversas.	*Margolis y Resnick* [MR00] Internet no es todavía un factor de transformación social.
Benkler [Ben06a] Internet genera una fuente auto-gestionada de información. Posibilita compartir conocimientos y experiencias, que hacen de cemento social para la esfera pública digital.	*Sunstein* [Sus01] En Internet se da una sobreabundancia de fuentes, con tendencia a la polarización, y en donde se reproduce los mismos discursos que en los mass media.
Savignys [Sav02] Internet como instrumento fuera del tradicional control de las élites. Además, ofrece una alternativa a la definición de la agenda[1] elaborado por los medios de comunicación de masas.	*Papacharissi* [Pap02] Internet ya se encuentra colonizada por los intereses comerciales.

> *lealtad* a las clásicas grandes instituciones (iglesia, partido o clase social) [Bec05, Ing97, Gid94]. Estas nuevas lealtades contribuyen a explicar por qué durante los últimos años en Estados Unidos aumentaba la intensidad en los denominados *lazos políticos débiles*, los de las asociaciones de voluntariado, por ejemplo, sobre los lazos de relaciones fuertes, tales como la familia o el trabajo [dlF10].

El el caso de España, el uso de las redes sociales se produce en función de los grupos etarios[1], con un perfil marcadamente generacional [ACGS10]. Los jóvenes encuentran un canal de comunicación para relacionarse, encontrar intereses compartidos, conformando una suerte de capital social digital [Put01].

En la arena política, han aparecido *espacios en disputa* denominados redes parapolíticas, que tienen una intencionalidad más o menos explícita, orientada a distintos tipos de acción política: difusión de noticias, reconstrucción de sucesos, identificación o adhesión simbólica a causas, o procesos de control y fiscalización ciudadana sobre la acción política de las instituciones, entre otros.

Las grandes organizaciones políticas, entre ellas los partidos, han penetrado en estas redes, tratando de influir en los comportamientos políticos. En muchas ocasiones, detrás de una demanda y su posterior politización, esto es, su entrada en la agenda política se encuentran los tanques de pensamiento o *think tanks* y grupos de presión organizados.

También se ha producido la aparición de *organizaciones híbridas*, gracias a la rápida adaptación institucional y la experimentación continua [Cha07a]. Estas redes siguen constituyendo un espacio por el que se compite, cuyo éxito vendría dado, en primer lugar, por la capacidad de construir redes y, segundo lugar, por poder interconectarlas con otras redes, a través de la cooperación en los objetivos y la combinación de recursos [Cas09].

El uso de distintas tecnologías de información y comunicación incluye un repertorio de actividades denominadas *tecnopolíticas* referidas específicamente al üso táctico y estratégico de las herramientas digitales para la organización, comunicación y acción colectiva como concepto clave para entenderlas. Esto supone la capacidad de las multitudes conectadas para crear y automodular la acción colectiva [Tor13]. En cambio, el término *ciberactivismo* [TQ12] tiene un carácter más amplio, en línea con Sampedro, quien ha definido distintos perfiles una tipología de personas según su comportamiento digital [Sam05, Sam11].

Denominados todos ellos como *internautas*, cada uno de esos tipos tiene una consideración que matiza su comportamiento en función del tipo de actividad que suelen realizar:

[1] Grupo etario es el grupo formado por personas de la misma edad.

- Los *internautas sociales* son quienes acceden a webs cívicas o de movimientos sociales.
- Los *internautas deliberantes* son los usuarios de blogs, foros y chats.
- Los *cibernautas* o *internautas partidistas* visitan webs de las candidaturas electorales con frecuencia. Dentro de este grupo, existen otros tres tipos según su grado de actividad. Los *ciberactivistas en general* son más proactivos en línea con la cultura de *Internet*, esto es, suelen intercambiar enviar mensajes o correos electrónicos con los partidos y candidatos. Los otros dos tipos están relacionados con su actividad dentro y fuera de los procesos electorales. Cuando un ciudadano desarrolla una actividad prolífica que tiene lugar durante las campañas electorales se habla de *cibernautas electorales*, mientras que se usa el término *cibernauta* no electoral o inactivo para referirse a los *cibernautas* que tienen una actividad notable, pero nunca durante la campaña electoral.

Existen diferencias del comportamiento de cada uno de estos tipos de perfiles. Así, su repertorio de actividades muestra una relación inversa al del compromiso que implica cada una de ellas: (i) consulta de información generalista en mayor grado que la estrictamente partidista o de movimientos sociales; (ii) envío de mensajes, de carácter impersonal y en muchas ocasiones en clave de humor; (iii) participación en foros y *chats*, en menor grado el envío de mensajes a partidos y candidatos; y (iv) firma de peticiones ante candidatos, partidos o instituciones.

Los internautas hacen uso de las nuevas tecnologías y canales para la expresión de opiniones, realización de propuestas y movilización política en gran medida porque las barreras de acceso -y por tanto el coste- de esos mecanismos de participación política son mucho menores en el nuevo escenario digital que a través de los esquemas clásicos de la comunicación tradicional.

La participación política y electoral de los internautas no se limita al espacio digital, sino que existe una relación significativa entre participación política en clave de reuniones y movilizaciones electorales tanto *online* como *offline*. De manera análoga, ciberactivistas sociales tienen una doble militancia, ya que toman parte en partidos políticos y movimientos sociales al mismo tiempo [LGVG11].

Con respecto a los internautas partidistas, éstos hacen un uso cada vez mayor de las webs de los partidos, a pesar del dominio de la televisión como canal preferido. Sin embargo, su grado de proactividad deja en evidencia el grado limitado de apertura que los partidos ofrecen en sus plataformas digitales fuera de sus agendas políticas.

2.1.2 **Los partidos políticos en Internet: usos de la web**

Los partidos políticos en España han incorporado progresivamente nuevos instrumentos, técnicas y estrategias de comunicación mediante tecnologías basadas en *Internet* desde finales de los años noventa del siglo XX.

Una de las líneas de trabajo más desarrolladas sobre esta cuestión es la de Dader *et. al* [DCC+14, DC08] que, basada en los modelos internacionales de análisis [GW00, FSX02, Ben06b], y como continuación al trabajo previo de [GD08], desarrolla un doble análisis cuantitativo y cualitativo del despliegue tecnológico en el sistema de partidos.

Desde el lado cuantitativo, de manera comparada, se tienen en consideración varias dimensiones, a saber: (i) la diversidad y profundidad de la información; (ii) el grado de interactividad; (iii) la sencillez de uso y ayuda instrumental y (iv) la estética y atractivo persuasivo. Por su parte, el enfoque cualitativo aborda un análisis de las funciones de (i) transparencia informativa de autopresentación, (ii) difusión y propagación informativa, (iii)función persuasivo-emitiva, (iv) función de movilización y reclutamiento, (v) función deliberativo-participativa, (vi) función de sencillez de uso y ayuda instrumental y, por último, (vii) la ubicación de los modelos elitista y participativo.

La primera gran experiencia una campaña electoral con gran despliegue digital tuvo lugar durante las elecciones generales de 2008. Durante esa campaña las webs de los partidos españoles durante ofrecen contenidos relacionados con la autopresentación, la disponibilidad de información de colectivos afines y otras instituciones externas; noticias y actualidad; información política de fondo (tal como el programa electoral o ciertos posicionamientos institucionales ante determinadas cuestiones relevantes o *issue*), información política de contexto. Así mismo, se realiza un esfuerzo en aspectos como la facilidad para interactuar en lectura simple, la posibilidad de respuesta a formularios y de contactar mediante correo electrónico, o bien la puesta en disposición de herramientas de expresión como foros o *chats*.

Se ofrecen instrumentos muy básicos orientados hacia la participación e interacción ciudadana, mientras que destacan las secciones de contenidos de actualidad mediante noticias elaboradas por los propios partidos. En cuanto a la movilización del electorado, se incluye información simbólico-comunicacional en cantidad suficiente, y numeroso material de propaganda política (con una proliferación de emblemas, eslóganes e incluso imágenes y fotografías, así como un diseño con cierta sofisticación y un cuidado del atractivo visual. Un solo

partido, UPyD[1], dispone de un canal para la captación de fondos.

En lo que respecta a los dos partidos principales del sistema en aquel momento, el PSOE alcanza una puntuación global superior, especialmente en el cuidado estético, la disponibilidad de información, tanto propia como de terceros, la interacción vía formularios, información de actualidad y puesta al día, la claridad de la estructura de contenidos así coma una profusa presencia de símbolos orientados a la movilización y la propaganda política. Por su parte, el PP obtiene una valoración con la media del resto de partidos, no destacando por ningún atributo en particular. Cabe destacar su grado de autopresentación y relaciones públicas, así como las facilidades para la participación mediante una notable oferta de foros de discusión y otras herramientas para a búsqueda de contenidos.

Desde el análisis cualitativo, existen dos modelos de referencia a los que se pueden adscribir los partidos políticos en cuanto a la relación con los demás:

a) *Modelo elitista (arriba-abajo)*. Ofrece una relación vertical entre el partido y otros actores políticos: en primer lugar, la relación entre el partido (su estructura interna) y su electorado; en segundo lugar, la relación entre el partido y la clase mediática (que representa a los medios de comunicación).

b) *Modelo participativo*. Este modelo muestra, por un lado, una relación más horizontal y en todo caso bidireccional entre el elector individual y el partido, y por otro lado, una relación muldidireccional entre el conjunto de electores con el partido.

En general, se constata una escasa información sobre el propio partido, incluidos los candidatos, responsables máximos y de nivel medio, en ambos partidos. En el caso del PSOE, resulta difícil el acceso a los perfiles biográficos de sus responsables, sin posibilidad de contactar mediante correo electrónico u otro mecanismo. En la misma dirección, no se hace uso de la potencialidad que ofrece la web, además de que los contenidos de alto impacto, como el programa electoral, se presentan de manera parcial y con acceso no sencillo. Con respecto al modelo de relación, es común en todos los partidos la falta de mecanismos estructurados para la participación efectiva. Con ello se manifiesta una clara orientación con el modelo elitista, según el cual los partidos están, bien orientados *hacia* los ciudadanos, no *con* los ciudadanos, o bien orientados hacia los medios de comunicación, de los que aspiran a fijar su agenda (con muchas noticias y documentos, generados a través de gabinetes de prensa de los propios partidos).

[1]UPyD: Unión Progreso y Democracia. Partido de ámbito estatal, constituido en 2007 y con representación parlamentaria desde las Elecciones Generales de 2008.

La participación en la mayoría de partidos se traduce en plataformas de adhesión en lugar de en espacios para la reflexión, la contraposición de argumentos o propuestas. El PP destaca por ofrecer una sección de foros que data del año 2000, y que muestra una relativa participación de ciudadanos con distintas posiciones ideológicas. Este servicio fue retirado de la web en 2009. Por parte del PSOE, la página personal del candidato, Rodríguez Zapatero, incluye un apartado con preguntas y respuestas, además de enlaces a otro tipo de comunidades *online*. En definitiva, aunque se declara oficialmente una apertura real a la participación digital a través de secciones con nombres como "Tu opinión cuenta.º "Participa", en la práctica no sucede así.

En las elecciones de 2008 todos los partidos ponen a disposición recursos en forma de blogs, información de los candidatos y los programas, pero de manera dispersa y sin una estructuración clara. La comunicación fue claramente unidireccional, a través del correo electrónico, *RSS* y boletines digitales a través de listas de correo. Esta comunicación tuvo un carácter claramente propagandístico, y muy poco expuesto al debate o la participación.

En cuanto a formatos de información, proliferaron los contenidos audiovisuales, ya fueran de elaboración propia o cortes de los emitidos por otros medios de comunicación. El uso de la comunicación emocional es notable, con una orientación hacia la persuasión emotiva. Con la excepción de UPyD no se hizo uso de la capacidad de captación de fondos a través de Internet, en un claro ejercicio de replicar los esquemas de financiación de Estados Unidos [Hin05], basados en *crowdfunding* o micromecenazgo.

2.1.3 La información política en España en la era digital

En el sistema político español, la principal fuente de información política la proporciona el sistema de medios de comunicación. En general, la información política es importante en relación a las actitudes y comportamientos políticos, además del puro conocimiento informacional que proporcionan. Sin embargo, son los partidos políticos los auténticos generadores de grandes volúmenes de contenidos ideológicos, programáticos y de acción política que, a través de los medios, llega a los ciudadanos.

Desde los años 80 se ha producido tres dinámicas en paralelo que han afectado al sistema de información y comunicación política: la proliferación de los canales de televisión privados, la desregulación en la programación en las radiotelevisiones, y una reducción y fragmentación en las audiencias y en los medios del Estado [CILSM09]. Estas audiencias, además, ya no se encuentran casualmente la información política, sino que la seleccionan de fuentes según sus in-

tereses. El consumo de información política ya no depende únicamente de que el contenido político se encuentre disponible, sino que se ve influenciado por aspectos como los hábitos, el interés o la competencia y ciertas características sociodemográficas.

Los efectos del consumo declarado de información sobre la participación política han sido objeto de debate académico, entre quienes mantienen que la exposición a información política lleva pareja una mayor tasa de participación, y quienes consideran que contribuyen precisamente a lo contrario: al cinismo y la desmovilización [Ben95b, RD90].

También ha recibido atención el efecto de la información sobre las campañas electorales, con posturas enfrentadas entre quienes afirman que el negativismo de muchas campañas conduce a una menor participación y la consiguiente desafección [AISV94, Put01, Pat02, CJ97]. Sin embargo otros académicos ofrecen una cara posibilista a dicha falta de afección, que en ciertas condiciones puede producir precisamente un mayor grado de movilización.

Existen dos aproximaciones a este tema. Mientras algunos estudios muestran una cierta relación entre consumo de medios y participación política [Lue09], otros plantean que existe una desigualdad de conocimiento político que en cierta medida está relacionada con las condiciones socioeconómicas de los ciudadanos [JSB00]. Tampoco existe consenso sobre si existe mayor implicación y participación entre quienes acceden a prensa escrita frente a quienes obtienen la información por radio, televisión o los nuevos medios. La evolución de las audiencias sigue los patrones expuestos por Shaw [Sha91], para quien en estos procesos de transformación de los sistemas de medios de comunicación, tras una fase emergente en un contexto de mentalidad abierta y tecnología innovadora se pasa a la etapa de adaptación, ajuste y búsqueda de nicho de mercado.

A partir de los datos del *Estudio General de Medios* (EGM) y los *Barómetros* del *Centro de Investigaciones Sociológicas* (CIS), los resultados de consumo agregado de medios tradicionales se han mantenido más o menos estables a nivel agregado, pero se constatan cambios a nivel individual [Mei10]. Estos datos están sujetos a ciertas restricciones de validez y fiabilidad, ya que se han encontrado evidencias de que este método de obtención de información presenta problemas de inflación, tal y como señalan Chaffee, Frank y Prior [CF96a, Pri05].

En el caso de la información política, se da la circunstancia de que, al mismo tiempo que se ha venido incrementando el número de ciudadanos que acceden a *Internet*, casi se ha cuadruplicado la cantidad de personas que consumen información política a través de este medio. Mientras tanto, para este tipo de contenidos, los medios tradicionales se han mantenido más estables, especialmente la prensa y algo menos la radio y la televisión. A ello han podido contribuir la

aparición de la prensa gratuita en el sistema de medios desde el año 2000 hasta el inicio de la crisis a finales de la primera década del siglo XXI. Se confirma, por tanto, que el desarrollo de tecnologías no conlleva necesariamente que se tenga más o mejor información política o que se consuma más [Mei10].

Mientras cabría suponer que las personas con mayor identificación partidista serían más proclives a un mayor consumo de información [Zal92], lo cierto es que esto sólo se produce en las personas declaradas de extrema izquierda. Existen claras diferencias entre quienes están habituados a este tipo de información frente a un grupo de personas a quienes no les interesa en absoluto. No se dispone de evidencias que nos indiquen si las fuentes preferidas para obtener información política a través de Internet son las versiones digitales de los medios tradicionales o si por el contrario, son otras nuevas fuentes presentes únicamente en Internet.

También ha quedado de manifiesto que el consumo de información política en cualquier medio –especialmente en la prensa– está relacionado con el consumo en otros medios, por lo que cobra sentido un perfil de personas con interés en la política y consumidor de información de esta naturaleza.

En una comparativa con el resto de países europeos en cuanto a consumo de medios, España estaría dentro del grupo de países del denominado *modelo mediterráneo o pluralista polarizado*. A partir de un desarrollo trabajo empírico, el marco teórico elaborado por Hallin y Mancini [HM04], plantea una tipología de sistemas de medios. Según esta clasificación, existen tres tipos de sistemas de medios, que vienen caracterizados en función de cuatro componentes principales, a saber: el desarrollo de los mercados de comunicación, y específicamente el campo de la prensa escrita masiva; la pluralidad política y social de dicha sociedad expresada a través de dichos medios; el grado de desarrollo de la profesión periodística en el país; y finalmente el grado de intervención estatal en el sistema de medios. Junto al modelo mediterráneo, en el que participarían la Europa del sur junto con Francia, están el modelo del norte y centro de Europa o corporativista democrático, y en tercer lugar el modelo atlántico o liberal.

Según este análisis, el modelo de medios español, en tanto que modelo mediterráneo, existe un notable grado de desigualdad entre el consumo de noticias en comparación con. Esta desigualdad se manifiesta en diferencias significativas en términos de sexo, con un consumo superior en los hombres. Por edades, el segmento de población de edad media (entre 35 y 49 años) es el que más consume radio y televisión, mientras que los ciudadanos con menores niveles de estudios leen la prensa en una proporción significativamente inferior [Mei10].

De nuevo, Hallin y Mancini señalan que algunas de las causas del modelo de medios están relacionadas con las circunstancias sociales, políticas e histó-

ricas del país, entre las que se encuentran el desarrollo tardío de la industria de la prensa, que inicialmente tenía como destinatarios a las elites políticas y económicas. Además de esto, el elevado grado de intervencionismo del Estado autoritario durante prolongados períodos de tiempo en los que no existían libertades democráticas propició condicionamientos e inhibiciones. Esto provocó que cuando se restauraron las democracias, los sistemas de medios -que nacen a partir de privatizaciones de los medios públicos- estuvieran fuertemente condicionados y controlados por elites de los partidos, la Iglesia o grupos sociales específicos. Estas desigualdades políticas llevaban parejas desigualdades sociales que en muchos casos se han mantenido a lo largo del tiempo.

Hay dos aproximaciones que tratan de explicar la desigual distribución de la información: la literatura del *knowledge gap*, y la del *telemalestar*.

- Para quienes postulan la existencia de un *knowledge gap*, los elementos clave son las distintas capacidades cognitivas y los factores sociodemográficos y sus consiguientes desiguales niveles de conocimiento [Gaz95, Kan08].

- Los trabajos del *telemalestar* nacen en Estados Unidos en la década de los setenta y ponen de manifiesto la gran influencia que la televisión ejerce como principal fuente para la obtención de información política. Este desajuste en el sistema de medios se traduce en una cierta contaminación en la cobertura de la información y su correlación con las actitudes y comportamientos de sus audiencias como ciudadanos. Los contenidos políticos estarían basados principalmente en los aspectos negativos de la política y la perspectiva sesgada de las campañas electorales como meras competiciones [Rob76].

  De esta manera, los efectos de la televisión se traducirían en un cambio de actitudes, especialmente la desconfianza y el desinterés por las cuestiones públicas y un notable incremento del cinismo unidos a una baja participación política y electoral. Como consecuencia de ello, los niveles de asociacionismo y participación política y electoral se verían mermados, traduciéndose en una reducción de los niveles de capital social en dichas sociedades [Put01]. La repetición de esquemas informativos en los que se presenta una imagen del mundo en el que triunfan la corrupción, el delito y el conflicto tiene consecuencias persuasivas en el medio y largo plazo [GGMS80].

  Una característica intrínseca de los nuevos canales de emisión de contenidos, el carácter audiovisual, tiene unas limitaciones específicas para la transmisión de informaciones políticas. Algunos autores señalan que en

> muchos casos, debido a su complejidad, la información política necesita ser explicada de manera más profunda. Sin embargo, lo que se produce es una simplificación y frivolización de los contenidos en aras a una supuesta facilidad de comprensión [Pos06, Sar98, Bor97].

Con todo ello, se producen varios fenómenos. El primero está relacionado con el incremento de la competencia en los sistemas de medios, que en España tuvo lugar en los años ochenta a partir del proceso de liberalización. A partir de este momento se observa un descenso en la atención en la elaboración de los materiales o contenidos, en lo que Millner denominó *tabloidización* de los medios [Mil02]. En segundo lugar, y como consecuencia de dicha competencia, las parrillas televisivas han tendido a ofrecer menos contenidos puramente informativos.

Los trabajos más recientes, a partir de un desarrollo empírico más robusto y, aceptando los planteamientos expresados en las líneas anteriores, tratan de matizar las consecuencias de los medios en los países europeos, que muestran cierta heterogeneidad en la naturaleza de las causas [HB90, New97, DVB06, dVBS06].

Pippa Norris ofrece una visión en torno a la idea de *efectos débiles* al correlacionar positivamente el consumo declarado de información formando un círculo virtuoso con el interés hacia la política y la participación política. Sin embargo, sucede que al mismo tiempo las personas menos interesadas, informadas e involucradas en cuestiones políticas son más susceptibles a la influencia de los medios, mientras que quienes sí están informados e involucrados consumen más información pero confían más en sus propias opiniones y creencia [Nor00].

2.1.4 Deliberación democrática y conversaciones *online*

Como se ha venido señalando anteriormente, en las sociedades democráticas se ha producido un incremento de la cantidad de ciudadanos que acceden al mundo *online* para involucrarse en la dinámica política [Bim03, Cha06], evidenciando que las redes de la sociedad civil se están ensanchando y al mismo tiempo, las distancias se vienen estrechando [Pax99, Put01, MSLB06].

La discusión política juega un papel crucial en el proceso democrático en la medida en que proporcionan la oportunidad de involucrarse en el proceso de conversación política y permiten formarse una opinión sobre las ideas opuestas [LBG65, Zuc05, Mut06]. Y es que estas conversaciones políticas pueden suponer las condiciones operativas para que se establezca una democracia deliberativa eficaz [Fre10, RMM$^+$10, CHLL09].

Existe cierto consenso científico de que, más allá de los individuos aisladamente, estas redes de discusión política constituyen un mecanismo de influencia social. Este efecto opera en dos direcciones. En primer lugar, las redes de discusión pueden ampliar las preferencias si los individuos interactúan con otras personas con ideas afines. Y en segundo lugar, pueden acercar posiciones y ayudar a elaborar consensos si abarcan diferentes rangos de opiniones [Sus07], lo que está relacionado con el *bridging capital*. Esto conecta directamente con la teoría de la democracia deliberativa, según la cual no todas las conversaciones constituyen una verdadera deliberación en tanto que para que la deliberación real tenga lugar se necesitan unas ciertas condiciones previas.

El modelo de González-Bailon et al., avanza en esta dirección analizando, no el contenido de las conversaciones online, sino la estructura de esas conversaciones [GBKB10]. A partir de los dos principios que sustentan la teoría deliberativa González-Bailon et al. plantean cuatro tipos o escenarios posibles. Esos dos principios son el principio procedimental, que describe el propio proceso de la discusión, y el principio de las consecuencias, esto es, el de los efectos de dichas discusiones.

Debido a que los ciudadanos a través de su interacción con otros ciudadanos conforman distintos tipos de red en función de la tecnología que usan, no todas las redes de conexiones cumplen la misma función permitiendo los mismos flujos de información. El modelo de González-Bailon et al. utiliza dos dimensiones. En primer lugar, considera las conversaciones en las que participan de manera voluntaria, y en segundo lugar analiza ciertas características de las discusiones. De este modo, incorpora las dos precondiciones para la deliberación: la representatividad de la discusión, y la presencia de esfuerzo persuasivo (medidos a través de la amplitud y la profundidad de las redes de discusión). Constata que las estructuras de las redes de discusión presentan diferencias significativas en función del tema sobre el que se esté debatiendo.

2.1.4.1 **La estructura de las redes de discusión: un modelo teórico**

El modelo propuesto por González-Bailon responde a dos cuestiones relevantes. La primera de ellas es si la estructura de todas las conversaciones es similar. La segunda se plantea que, en el caso de no ser así, cuál de ellas se aproxima más al modelo deliberativo ideal.

Se caracteriza a las discusiones a través de dos rasgos estructurales básicos: su anchura y su profundidad. La *anchura* mide el número máximo de comentarios en cualquier nivel de la red; la *profundidad* mide el número de capas o niveles que despliega la discusión. Sirven para aproximarse al número de perso-

nas diferentes que participan y también para conocer la intensidad de la argumentación. Cruzando ambas posibilidades, según este modelo teórico, existen 4 tipos ideales de conversación:

- *Tipo I*: Atraen la atención de un gran número de personas a la conversación, en la que muestran una gran intensidad en sus interacciones.
- *Tipo II*: Aunque existe una gran cantidad de interacción, en realidad son muy pocas las personas que participan en estas conversaciones.
- *Tipo III*: En estas conversaciones no existen muchas personas distintas que participen, y cuando lo interactúan tampoco lo hacen con poca intensidad.
- *Tipo IV*: Son conversaciones en las que se da una elevada concurrencia de participantes, pero ofrecen una argumentación pobre en intensidad.

En sus investigaciones, trabajan con dos modelos de análisis, un modelo básico y un modelo avanzado:

- **Modelo básico**. Este modelo considera solo el número máximo de comentarios en cada nivel de la red, y el número de capas o niveles anidados. Con ello se diferencian distintos tipos de discusiones en función de las dinámicas colectivas que generan y, por el otro, identifican dónde las discusiones políticas caen en el vacío. Mediante una definición limitada de "anchura" y de "profundidad" sobre 2 modelos, se constata que en foros de discusión de carácter general, los datos de todas conversaciones muestran que sólo un 5.5 % de ellas son conversaciones políticas.

  La discusiones varían significativamente en cuanto al tipo de dinámicas que generan: algunas es más probable que atraigan a un número mayor de participantes, mientras que otras tienden a incitar mayores cadenas intercambio entre los participantes en la discusión. Estas diferencias dependen del tema, de modo que en discusiones sobre asuntos políticamente relevantes convocan a un mayor conjunto de participantes y se involucran en interacciones más intensas.

- **Modelo refinado** A partir del modelo básico, incluye variables adicionales de control para la presencia de participantes prolíficos y para los efectos de los comentarios particularmente conflictivos en la estructura completa. Para ello, añade 4 variables: (i) el número máximo de comentarios en cualquiera de los niveles; (ii) como medida de amplitud incorpora el número total de comentarios y el número total de usuarios únicos en la

discusión; (iii) como medida de profundidad utiliza dos versiones del *h-hindex* científico[1]. Los resultados muestran que la conversación política *(political talk)* tiene una cualidad intrínseca que la diferencia de otros tipos de debates. Y además, evidencia que no todos los tipos de discusión pública conlleva las mismas dinámicas colectivas.

A partir de ambos modelos se formaliza una base teórica elaborada a partir de la definición de un modelo empírico que permite comenzar a diferenciar la deliberación de la mera discusión. En segundo lugar, se propone un marco de trabajo a partir del cual abordar análisis comparativos y analizar distintas configuraciones *online* a la luz de las dinámicas deliberativas que promueven las conversaciones políticas. Además, esta aproximación permite clasificar una determinada discusión cuantificando hasta qué punto se aproxima o no a un determinado ideal-tipo.

Una vez realizada una aproximación a los cambios producidos por la aparición de Internet en la política, en la sección siguiente vamos a conocer las principales claves que pueden explicar la participación electoral y, en particular, los modelos teóricos contrastados en los estudios electorales centrados en el voto.

2.2 Estrategia y competición electoral

Tras describir los principales cambios que los avances tecnológicos han venido produciendo en los últimos años en el ámbito de la política, en esta sección se describen aquellos aspectos relacionados con el acceso al poder por parte de los partidos políticos. En particular, se señalan los procesos electorales como un momento en el que se ponen en acción gran parte de las energías de los partidos.

Los partidos políticos son instituciones con una historia reciente, ya que apenas han existido desde el siglo XIX a partir de los movimientos sociales de clase. Han sido un agentes clave en el desarrollo de la democracia liberal y como tales desempeñan una función esencial de representatividad de intereses colectivos y, a partir de ello, desempeñan otra función relacionada con el gobierno y el ejercicio del poder político. Actúan en las sociedades dentro de los sistemas políticos y por lo tanto producen entradas o *inputs políticos* que el sistema convierte en salidas o *outputs políticos* [Eas53]. Estos *inputs* y *outputs* están formados por elementos de diversa índole. Así, se distingue entre elementos cognitivos y otros

[1] El *h-index* es un indicador que mide la calidad profesional de físicos y de otros científicos, en función de la cantidad de citas que han recibido sus artículos científicos. Así, se dice que un científico tiene índice h si ha publicado h trabajos con al menos h citas cada uno.

elementos psicológicos tales como juicios, valores y afectos, con respecto a los comportamientos electorales (la participación en un determinado mítin, o el ejercicio del voto entre muchos otros).

Durante las décadas de los 40 y 50, bajo el influjo de las corrientes conductuales o *behaviouristas* se plantearon las campañas como una técnica que pretendía, a partir de la observación del comportamiento de los votantes, tratar de influirles hacia su voto.

En las dos décadas siguientes, esta perspectiva cambia por otra que considera al elector como un sujeto que está inmerso en una sociedad. Inaugura el enfoque "sistemático-institucional", que además de considerar los elementos culturales, incorpora los elementos institucionales de una determinada sociedad, a saber, el sistema de partidos o el sistema electoral entre otros. El refuerzo de esta visión cultural del ciudadano-elector se incrementa durante la parte final del siglo XX, aglutinando tanto elementos de carácter social, como elementos psicológicos racionales y emocionales [Can98].

En todo caso, las campañas tienen un enfoque claramente persuasivo, esto es, dirigir al ciudadano-elector hacia el voto a una determinada opción política. En consonancia con esto, la comunicación de campaña se caracteriza por: (i) un contacto directo entre candidatos y electores en mítines y *rallies* en Estados Unidos; (ii) la deliberación en forma de discurso y debate, de manera que se trata de convencer al elector de una u otra opción; (iii) los recursos utilizados en campaña toman forma de folletos, vallas publicitarias, o los programas electorales en soporte escrito; (iv) se realiza un envío directo de material de campaña a través de correo.

En paralelo a estos desarrollos, ha cobrado una vital importancia el uso generalizado de medios de comunicación a través de los cuales trasladar la información política desde los partidos hacia los ciudadanos, al mismo tiempo que del lado de los propios partidos la gestión de la comunicación se ha ido profesionalizando a través de profesionales expertos, en muchos casos provenientes de medios de comunicación.

2.2.1 **Los ciudadanos ante el voto: perspectivas y modelos de comportamiento electoral**

Si el propósito electoral de los partidos es obtener representación parlamentaria y estar en disposición de formar gobierno, resulta necesario comprender los fundamentos del sistema electoral y del comportamiento político, en especial, del voto. Los estudios sobre la relación entre los sistemas electorales y los sistemas

de partidos parten de los trabajos clásicos de [LA94, Rae67] y han ido incorporando las principales transformaciones del lado de los partidos políticos, como los recientes trabajos de Lawson [LM14] o Inglehart [Ing15].

A la hora de conocer el comportamiento político electoral, tres son los grandes enfoques desde los que se ha abordado históricamente el voto:

1. **Modelo sociológico del voto**
   El modelo sociológico del voto tiene su origen en tres trabajos seminales: "*The People's Choice*" [LBG65], "*Voting*" [Ber86], y "*Personal Influence*" [LK55] en los que se utilizó por primera vez el cuestionario como técnica de investigación social. La tesis de Lazarsfeld era que el acto de votar es un acto individual, que se ve influenciado por la personalidad del votante y su exposición a los medios. Sin embargo, los resultados mostraron que el efecto de los medios es mínimo, y que es el contexto de los grupos sociales lo que juega un papel clave.
   A este respecto, los principales factores que explicarían el voto de los electores son: (a) la homogeneidad política de los grupos sociales, (b) la activación, refuerzo y conversión, (c) la transmisión social y las opciones políticas, (d) la percepción y diferenciación política, (e) la teoría y la práctica democráticas, y (f) el anclaje o *clivaje* social.

   (a) **Homogeneidad política de los grupos sociales**. Según Lazarfeld et al. averiguaron, la mayoría de los electores votaban según su predisposición política original, que estaba descrita a través del estatus socioeconómico, la religión y el área de residencia. El cambio en el sentido del voto se vio que estaba influenciado principalmente por seguir las "presiones" de los miembros de su comunidad de pertenencia [LBG65].

   (b) **Activación, refuerzo y conversión**. Del estudio de las campañas electorales y los procesos de comunicación se producían tres tipos de efectos: (i) la activación de los indiferentes, (ii) el refuerzo del vínculo de quienes estaban próximos a un partido, y (iii) la conversión de los indecisos. Los resultados mostraron un efecto menor de las campañas electorales. Ahora bien, lo que sí producían era una cierta cristalización de la cohesión de los simpatizantes o *supporters*, más que un cambio de opinión. Esta *atención selectiva* de votantes era precisamente el objetivo que los jefes de campaña pretendían hacer cambiar.

   Sí se constató que los grupos de pertenencia resultaban cruciales, especialmente las personas más relevantes de dichos grupos, a la hora

de filtrar la información y la opinión que los miembros del grupo recibían de los medios de comunicación. Estos "líderes de opinión", concepto que aparece en el capítulo *"The Nature of Personal Influence"*, desarrollado más tarde por Katz y Lazarfeld en *"Personal Influence, The part played by people in the flow of mass communications"* [KL70] conformaban el denominado modelo de comunicación en dos fases: (i) de los medios de comunicación hacia ellos, y (ii) desde ellos hacia el resto de los miembros de esa comunidad. Estos líderes elaboran y reelaboran permanentemente estos contenidos en sus círculos sociales. Se encontró una importante influencia entre la relación entre los grupos sociales de pertenencia, la opción política y el papel decisivo de los contactos personales a la hora de construir el voto, de tal modo que estos procesos de decisión electoral son procesos de grupo de cohesión, más que individuales.

(c) **Transmisión social y opciones políticas**. Ampliando el estudio anterior a las presidenciales de 1948, los resultados que confirmaban los anteriores. En primer lugar, la "diferenciación social", basada en el estatus socio-económico, religión, raza y lugar de residencia, es una precondición para la disensión política. En segundo lugar, existen tres condiciones para la transmisibilidad que asegura el mantenimiento de esta diferenciación de generación en generación, y finalmente, las condiciones para una proximidad social y psicológica entre miembros de un grupo facilita y mantiene el *clivaje* electoral. Estos tres procesos (diferenciación, transmisión y contacto) aseguran la transmisión y la elección política (*political choice*).
Además de esto, la diferenciación social marca una inicial natural posición política alrededor de la del grupo de pertenencia, que es sobre la que pivotarían posteriores procesos de reactivación [Ber86].

(d) **Percepción y diferenciación política**. Dada la importancia de las cuestiones económicas, se producía la circunstancia de que, solo la mitad de los votantes de cada partido apoyaba las medidas económicas de su partido. Esta disonancia se resolvía, no cambiando el sentido del voto, sino modificando la percepción del propio candidato.

(e) **Teoría y práctica democráticas**. De este modo (con los 2 trabajos citados), llegan a la conclusión de que el interés, la discusión y la motivación, el conocimiento-información, los principios y una baselógica para decidir el voto son elementos sustanciales en el comportamiento político. Sin embargo, la mayoría de los votantes no tienen

ningún interés en las cuestiones relacionadas con la política.

(f) **Clivaje social** Este enfoque procede principalmente de la Escuela Sociológica, cuyo referente es la Universidad de Columbia. Propone una explicación del voto relacionada con la estructura social. Según este enfoque, se han formulado *clivajes* (del inglés *cleveage*) o anclajes distintos que han venido operando con efectividad en distintas épocas y sociedades. Así, se habla de clivaje social, clivaje religioso, clivaje de origen (ya sea nacionalidad o étnico, el centro-periferia o el rural-urbano) que están relacionados con una determinada expresión del voto. Cada uno de estos clivajes marca un punto de anclaje o división para una parte de una comunidad o sociedad, de modo que cada persona se autoasigna de manera binaria a cada lado de dicha división. La aproximación micro-sociológica de Lazarsfeld y Berelson les permitió identificar ciertos procesos de naturaleza psicológica tales como la percepción persona, la identidad social o las preferencias intra-grupo. Estas formulaciones permitieron el desarrollo posterior de las aproximaciones psicológicas [LBG65, Ber86].

Los posteriores desarrollos del modelo sociológico se centran principalmente en el estudio de los sistemas de partidos (*"Political Man"* de Lipset 1960 [Lip63], "*Party Systems and Voter Alignment*" de Lipset y Rokkan 1967) [LR67]. Estos sistemas de partidos se van consolidando en los países occidentales debido a procesos de revolución nacional (que responden los anclajes centro-periferia y estado-iglesia) o bien a la revolución industrial (en los que los anclajes más eficaces con el urbano-rural y capital-trabajo) [MB99].

Aunque el modelo sociológico no consigue explicar del todo el comportamiento en el voto, sí consigue al menos dar una explicación del voto en términos de comportamiento estable a largo plazo. En el corto plazo, las matizaciones vendrían explicadas en términos de lo que el modelo psicológico de Michigan denomina "identificación partidista" o *partisanship*, y que trata de ligar las influencias sociológicas de largo plazo y los factores sociales y políticos en el corto plazo de cada elección.

2. **Modelo psicológico del voto**
Desarrollado por la Escuela de Michigan, este paradigma señala que son las actitudes, los valores y la ideología los que marcan la orientación del voto. La socialización es, en este caso, un proceso clave, que contribuye de manera notable a explicar variables como la identificación partidista. Sobre esta identificación subjetiva se construye el *Modelo Ideológico del Voto*

que ha explicado cómo se ha votado en distintas elecciones en muchos países y momentos de la historia. Según esta Escuela, existen otros dos tipos de identificaciones subjetivas, que son la identificación izquierda-derecha o la identificación nacionalista, y que influyen sobre la identificación partidista. Tiene como origen los estudios de las elecciones presidenciales de 1948, 1952 y 1956, analizados por Campbel y Kahn en *"The People Elect a President"* [CK52], *"The Voter Decides*" (1954) [CGM54] y *"The American Voter"* [Cam80]. Según esta escuela, los principales factores que explican el voto serían: la identificación partidista, el *funnel* o túnel de causalidad, la proximidad o la distancia ideológica.

(a) **Identificación partidista**. Según este modelo, existe un concepto central, la *identificación partidista* (del inglés *partisanship*), que muestra una afinidad psicológica, una relación estable y duradera hacia un partido, lo que no necesariamente se materializa en un vínculo concreto, afiliación o siquiera un voto permanente o una militancia sistemática con dicho partido.

El concepto de identificación partidista se vio influenciado por el concepto de "grupo de referencia"[HS68], de modo que esta identificación emocional se adquiere a través de un proceso de socialización, influenciado por los valores y actitudes de la familia, los amigos y los pares, adquiriendo distintos grados de implicación (Miller y Shanks 1966). Sin embargo, el modelo no expone una relación directa-causal entre esta componente subjetiva (*partisanship*)y la componente objetiva del voto, sino que propone que aquella actúa como un filtro a la hora de percibir lo que los votantes aprecian que es favorable a la orientación de dicho partido e ignorar o devaluar lo que se opone a él.

(b) **Funnel de causalidad**. El modelo del *túnel de causalidad* representa una cadena de factores y eventos que afectan a los subsiguientes dentro de una secuencia que finalmente conduce a que el ciudadano elabora su voto. Así, existen tres grupos de factores básicos (factores socioeconómicos e históricos, valores y actitudes, y grupos de pertenencia), a los que se añaden otros de alcance más próximo o inmediato (*issues* o temas específicos, el candidato, la campaña electoral, la situación política y económica, la acción de gobierno y la influencia de amigos). Este modelo pone el acento especialmente en la identificación partidista, los candidatos y los *issues*, y menos en los aspectos sociales y en los sistemas de comunicación [NWK84].

(c) **Distancia ideológica**. Aunque el modelo del túnel plantea que pueden producirse cambios a largo plazo en la identificación partidista, para Campbell se trata de algo absolutamente raro, únicamente como reacción a sucesos con un gran impacto. Estos cambios se manifestarían a través de actitudes ante propuestas políticas, candidatos y beneficios al grupo [HM87].

Desde su formulación en los años 60, el papel del *partisanship* ha sido cuestionado, especialmente por dos motivos: la dificultad de aplicarlo a países con sistemas de partidos no bipartidistas [Bud76], así como por su falta de eficacia ante la progresiva desafección hacia los partidos políticos en los Estados Unidos [NP99, SBN86, SN88] y también en Europa [CD85, Dal84, DFBA84, SH95]. Todos estos fenómenos ponen en cuestión la fiabilidad del *partisanship*, de modo que cuestionan una cierta sobrevaloración a la hora de describir su papel en las lealtades de partido. En cambio, se propone que, en cada elección, los votantes evalúan las circunstancias económicas, y cómo los partidos y sus líderes las gestionan. De este modo, el *partisanship* actúa, como mucho, como un simple atajo cognitivo y no como una especie de dispositivo de identidad social que ayuda a elaborar una decisión electoral. Se propone que a la hora de realizar esta evaluación debe de existir un mecanismo racional que permita procesar la información.

3. **La elección racional o *rational choice*** Basado en los principios económicos que Downs describe en *"An Economic Theory of Democraccy"* [Dow57], bajo la premisa de que equiparar los parámetros económicos (recursos, mercancías y tecnología) con los resultados políticos. En este sentido, existe una clara analogía entre consumidores y votantes, y entre empresas y partidos políticos. Para ello hace uso de tres premisas, la primera de las cuales consiste en que las decisiones de partidos y votantes son racionales; en segundo lugar, los sistemas políticos democráticos asumen un cierto nivel de consistencia de modo que admite predicciones sobre las consecuencias de partidos y votantes; finalmente, para que exista un determinado nivel de oferta electoral es necesario que los sistemas dispongan de cierto margen de incertidumbre que posibilite una diferencia en dicha oferta.

   Este compendio teórico está representada por cuatro modelos de explicación del voto:

   - En primer lugar, el sobre *Modelo de Competencia Espacial* en el eje izquierda-derecha se plantea la pugna por el centro político.

- Un segundo modelo propone influir mediante campañas electorales y los medios de comunicación. Se sustenta en la teoría hipodérmica, según la cual el receptor (que es el elector) recibe una comunicación política en campaña, la asume, la analiza y le influye a la hora de votar.
- El tercer modelo es el del voto temático (*issue voting* en inglés) plantea que todos los electores tienen una opinión determinada sobre cualquier tema y en especial sobre temas cuyunturales, el ciudadano está interesado sobre dichos asuntos y además los electores conocen la posición de los partidos sobre dicho temas.
- Por último, según el modelo del voto económico, se ejercerá el voto en base a los resultados económicos que proporciona al individuo el partido que gobierne. Este voto económico puede plantearse en términos individuales o personales, o bien en términos colectivos o sociales.

(a) **Maximización de la utilidad de las acciones**. Según Dawns, el concepto de racionalidad se construye a partir de la asunción de que tanto votantes como partidos actúan directamente según sus propios intereses. Se trataría de una perspectiva de utilidad privada que actuaría minimizando el uso de recursos para maximizar un resultado individual. En el caso de los partidos políticos, definen sus estrategias con el único objetivo de ganar las elecciones y llegar al poder, de tal suerte que formulan políticas para ganar las elecciones, y no al revés, ganar las elecciones para formular políticas.

(b) **Consistencia**. Para una evaluación racional, un votante necesita poder hacer predicciones y poder compararlas, en una especie de análisis comparado o *benchmarking* de valor calculado esperado para cada una de las alternativas. Por este motivo es necesaria cierta consistencia en el comportamiento de los partidos si están en el gobierno, aunque esto resulta mucho más complicado si no es así.

(c) **Incertidumbre**. Aunque se podría esperar que cuando un partido llega al gobierno trate de liderar y defender propuestas a favor de los intereses de la mayoría, esto casi nunca sucede así. Los partidos políticos presentan propuestas suficientemente diferenciadas a un electorado para movilizarlo de cara a que participen votando. Por tanto, la oferta se basa en último término en producir esa movilización, por lo que existe una mezcla de ideología y ambigüedad en los discursos de los partidos. Es por esto que los partidos tienden a definir

una ideología de cara a maximizar los apoyos del mayor número de votantes. Estos ajustes dinámicos están limitados por la consistencia citada, que además debe ser predecible de cara a un hipotético comportamiento futuro. Según la teoría de la elección racional, lo que importa a los votantes no es la ideología sino las acciones concretas que toman los gobiernos.

(d) **Representación espacial**. La posición relativa de partidos y votantes se representa en un modelo unidimensional lineal, numerado de izquierda a derecha desde cero a cien. Los votantes tratarán de elegir a aquellos partidos que más cerca estén de su posición, mientras que por su parte los partidos se aproximarán al punto en el que más votantes existan.

Cada uno de estos tres grandes enfoques se han abordado desde dos perspectivas: el análisis macroagregado o ecológico, y desde los análisis individualistas o *cross-level*. Todos ellos señalan que la participación electoral es clave, cuyos factores pueden tener carácter individual o contextual:

- *Factores individuales*: sociodemográficos (edad y/o ciclo generacional, nivel de estudios, nivel de ingresos), actitudes y valores (interés por la política, politización o identificación política subjetiva, satisfacción política entre otros).

- *Factores contextuales*: el contexto institucional (el ámbito de las elecciones, el sistema electoral), el contexto partidista (número de partidos, grado de polarización, nivel de competitividad [Dow57], los anclajes, la capacidad de movilización de los partidos, etc.) o los incentivos a participar.

La comunicación, como elemento cosustancial al ser humano, ha cobrado una enorme importancia en las sociedades modernas, y los partidos políticos han recogido gran parte de esta experiencia y conocimiento, principalmente del ámbito de la gestión empresarial.

Con la democracia liberal a partir del siglo XIX, los movimientos sociales y después los partidos políticos han sido conscientes de la enorme importancia que la comunicación política tiene en el camino hacia el acceso al poder político y también una vez que se está en ejercicio del gobierno.

En especial, la comunicación política ha vivido una gran auge, especialmente desde la explosión de los grandes medios de comunicación social: primero el periódico y los medios impresos, posteriormente la radio y después la televisión. Durante las últimas décadas, con el despliegue de Internet se está viviendo una

auténtica revolución que está transformando no solo la comunicación política, sino la forma en que los ciudadanos se relacionan entre sí y con las instituciones políticas y los partidos.

Las campañas electorales se presenta como un escenario idóneo en el que los partidos políticos despliegan estrategias y tácticas comunicativas cada vez más sofisticadas. En las páginas siguientes se describen los principales elementos del análisis electoral así como el papel que desempeñan las campañas electorales.

2.2.2 Las campañas en el proceso electoral

En las democracias liberales, las elecciones son un instrumento democrático con un notable desarrollo teórico y práctico en los últimos siglos. Los sistemas electorales tienen una función principal, que es la de proporcionar representantes institucionales a los ciudadanos. Estos esquemas de representación , a su vez, desempeñan una misión de orden superior: formar gobierno y proporcionar legitimidad a los representantes del sujeto político que los elige [LA95, Sar80, Lin87, Noh13].

En estos procesos las campañas juegan un papel importante en la medida en que sirven para que las opciones políticas informen de sus oferta programática a los ciudadanos, incorporen propuestas y mejoras a dicha oferta y consigan finalmente una mayoría que les de acceso a las instituciones representativas y de gobierno.

De cara a dicha representación, los partidos políticos compiten para alcanzar el poder. Actúan proponiendo una oferta que es seleccionada y priorizada por el sujeto electoral. Por tanto, esta oferta ha de ser conocida para poder ser posteriormente valorada por los votantes. La comunicación política electoral desempeña esta función, que cobra notable interés durante los procesos electorales, pero que no se circunscribe únicamente a dichos momentos.

Las campañas electorales han evolucionado sustancialmente desde las primeras épocas hasta la actualidad, debido a razones de diversa naturaleza, pero en general comparten algunas propiedades que se mencionan:

- Dar a conocer las propuestas programáticas e ideológicas de cada uno de los partidos que se postulan a obtener representación institucional, permitiendo confrontar dichas posiciones a través de mecanismos como los debates electorales entre candidatos.

- Sirven para canalizar el enfrentamiento político y lo resuelven pacíficamente, permitiendo gestionar cambio y continuidad en un proceso estruc-

turado y normativizado, esto es, recogido en normas.

- Permiten fijar los asuntos colectivamente prioritarios; informan a los votantes; y muchas veces son determinantes en el resultado electoral cuando están muy igualados.

- Son instrumentos imprescindibles en la medida en que durante muy poco tiempo condensan un fuerte contenido simbólico. Atraen la atención de los ciudadanos sobre ciertos asuntos que se consideran relevantes y concitan el interés sobre el carácter de los candidatos.

- Las campañas permiten asentar predisposiciones favorables a la audiencia y mitigar las desfavorables.

Durante los últimos años, las campañas electorales, al igual que gran parte de la práctica de la comunicación política, ha venido adquiriendo algunas propiedades que las definen.

- *Campaña permanente*.En muchos países, la práctica de los partidos y del resto de actores del sistema político ha conducido a una campaña perenne durante todo el período del mandato de las instituciones. En tal sentido se interpreta el concepto de *campaña permanente* [OM00].

- "*Espectacularización*"de las campañas. En las últimas décadas se han incorporado elementos de otras artes y disciplinas, como el cine, la televisión, añadiendo componentes de la publicidad y la comunicación empresarial y el entretenimiento [GCHS92, SW94a, Car12]. La campaña de Obama en 2008 tiene una gran componente de dominio de la retórica, evidenciando el poder de la escenificación, el arte del relato y las nuevas tecnologías digitales. La experiencia política cede el lugar a la "competencia ficcional", para instalarse en la "representación narrativa" [Arr12].

- Los mítines constituyen uno de los elementos clásicos de las campañas electorales. Suponen unos rituales muy parecidos en todo el mundo, y aunque la tecnificación ha tenido grandes consecuencias con respecto a los efectos de los mítines, el elemento singular continúan siendo el discurso y el candidato. En la representación del discurso, la lógica "nosotros-ellos" continúa estando en vigor.

- *Campañas mediáticas*. El uso de medios de comunicación para la distribución de las propuestas. Radio, televisión, actos y eventos con electores, simpatizantes. Los debates, y el debate sobre el debate (de ahí aparece

la figura del *spin doctor* que trata de revertir la interpretación del resultado de un debate político entre candidatos).En cuanto a publicidad, los anuncios electorales, varios subgéneros: cambio, solvencia y seguridad, esperanza, el pueblo frente a los poderosos, identidad nacional, parodia del adversario, amenaza [HH10]. Las campañas electorales en Estados Unidos se han venido definiendo como *post-modernas*, y se caracterizan por un uso masivo de la televisión, un electorado pasivo y una lógica propagandística [Nor00].

- *Campañas personalizadas en candidatos*. Frente a las campañas orientadas a ideología o programas), las nuevas campañas electorales muestran un perfil elevado de los candidatos. Con respecto a los candidatos, la posición relativa de sus opciones para alcanzar representación influye sobre su comportamiento a lo largo de la campaña. Así, los candidatos favoritos arriesgan poco y los desafiantes arriesgan más. Del mismo modo, el carácter de los candidatos se convierte en centro de atención: su templanza, su cercanía, su voluntad de servicio.

- *Contenidos pagados*. Desde los años setenta, los anuncios pagados en televisión suponen el mayor gasto de las campañas en países como Estados Unidos o Reino Unido, donde la publicidad política es un campo altamente desregulado, lo que ha motivado que las candidaturas con mayor financiación sean las que más puedan invertir en este tipo de formato [Wes13].

- *Elevado grado de profesionalización*. La profesionalización de las campañas ha venido de la mano de la proliferación de encuestas permanentes, expertos en marketing, escritores de discursos. Al mismo tiempo, las dosis de información de campaña se empaquetaban y distribuían de modo masivo a través de los medios principales [AC08].

Las campañas electorales para las Elecciones Generales de 2008 en España se caracterizaron por una estrategia de los partidos basada en la campaña permanente, con un elevado contenido de campaña negativa e intensamente personalizada en las figuras de los candidatos. Las herramientas tecnológicas fueron utilizadas de manera distinta a como se habían utilizado antes (ligadas a los movimientos sociales), sino como auténticos instrumentos de marketing, principalmente a través de los canales que los dos grandes partidos habían creado en sitios sociales como *YouTube*, *Facebook* e incluso *SecondLife*[1] [LGVG11, SP08]. El espacio de una nueva política 2.0 en la que *Internet* actúa como espacio para

[1] `http://www.secondlife.com`

la reflexión se truncó por el uso partidista de estos canales, aumentado por la reproducción de sus mensajes a través de los canales digitales de los medios de comunicación. El resultado de una clara estrategia *top-down*, a diferencia de la de Obama unos meses después.

Recordando la tipología de perfiles en Internet de Sampedro, hemos de destacar describe tres tipos de perfiles en función del tipo de actividad que llevan a cabo en la red. En primer lugar, los *ciberactivistas sin fundamento* reenvían información sin analizar o reflexionar sobre su contenido (a veces sin consultarlo) o sus efectos. Este tipo de comportamiento supone un ejercicio de banalización de la ciberpolítica. En segundo lugar, los *ciberactivistas lúdicos* generan y reenvían gran cantidad de contenido de *infosátira* (esto es, que transmite información desde la burla o sátira), junto a parodias de los candidatos y de los partidos, sobre propuestas partidarias a destinatarios con cierta afinidad ideológica. En tercer lugar, los ciberactivistas molestones o *trolls* publican y transmiten contenidos lúdicos o con distinto grado de provocación o molestia. Es la figura del *trol*, ya estudiado por [TSFW05, Har10, DlP14]. Este rol, desarrollado en el ámbito de Internet, pretende introducirse en conversaciones de terceros con el propósito de interrumpir el fluir natural del grupo. Por último, existe otro rol desempeñado por personas que solo quieren manifestar su opinión en un blog o en un foro. Principalmente son jóvenes y no existen evidencias claras de que tengan una gran componente de compromiso cívico y/o político. Este es el rol de los ciberactivistas expresivos.

En cuanto a la pertenencia a algún tipo de organización, se distinguen distintos perfiles de ciberactivistas:

- Los *ciberactivistas implicados* promueven su afinidad a un partido o ideología hacia otras personas. Aunque no forman parte de la maquinaria de ningún partido, propagan esta afinidad reenviando información o promoviendo la movilización política (tal como asistir a una manifestación o mítin) y también elaborando material audiovisual a distribuir en sus mensajes.

- Los *ciberactivistas de partido* forman parte de las estructuras de los partidos, y dedican parte de su jornada a elaborar contenidos, publicar mensajes y reenviar esta información a través de canales como el correo electrónico, los blogs, los foros o las redes sociales. Al servicio de sus organizaciones, trabajan sobre con una línea editorial en forma de argumentario.

- Los *ciberactivistas sociales* están vinculados a movimientos ciudadanos y movimientos sociales, por lo que tratan sobre distintos temas, manteniendo posiciones distintas sobre temas de sociedad y política. Aunque tienen

no pertenecen estrictamente a la lógica de partidos, sí tienen relación con las posiciones ideológicas de éstos, por lo que han de considerarse.

Todos ellos comparten una visión similar respecto de Internet, que les proporciona un medio con mayor autonomía, un grado menor de estructuración -y por tanto menos restrictivo-, que permite niveles mayores de interacción y que hace posible la proactividad de los participantes. En el mismo sentido, se entiende que Internet ofrece un modelo de comunicación y acción política más dinámico, frente a la acción política más convencional, caracterizada por plantearse de arriba abajo, , orientado al marketing político. Se entiende Internet, no como un mundo aparte, sino como una parte más del mundo.

En España se constatan algunas adversidades al modelo de ciberpolítica:

- El carácter *virtual* de la red: impacto reducido, penetración limitada, sesgo y barrera generacional, pobre cultura participativa.
- La ausencia de un discurso propio sobre la ciberpolítica, más allá de la función meramente simbólica de las TICs, las reducidas consecuencias, la creciente espectacularización y negativismo de la ciberpolítica (en línea con el resto de la campaña 2008).
- Los discursos más elaborados y profundos se dan a medida que crece el grado de ciberactivismo y la identificación ideológica.
- Una crítica más sofisticada a los argumentos de los demás en función directa, al igual que antes, del grado de uso y del conocimiento político.

Hasta este momento hemos descrito el impacto que Internet ha producido en la arena política, en determinados comportamientos sociales y sus efectos en el consumo de información política. En segundo lugar, hemos realizado un recorrido por las principales escuelas de pensamiento y sus modelos a la hora de explicar el voto. A partir de este momento abordaremos el tercer vértice que constituye el área de nuestra investigación, la comunicación política.

2.3 Comunicación para la competición electoral

En esta sección nos acercamos a la comunicación política desde tres elementos centrales en nuestra investigación. En primer lugar, abordamos una caracterización de las campañas electorales y su dimensión en la esfera digital, haciendo especial énfasis en la deliberación democrática. En segundo lugar, se repasa

la comunicación electoral en la red social *Facebook*. Y cerramos la sección exponiendo las teorías de análisis de marcos comunicacionales, destacando los aspectos que tocan con el ámbito electoral.

La comunicación política es una actividad de la política que implica el intercambio de símbolos o mensajes que con un significativo alcance han sido compartidos por, o tienen consecuencias para, el funcionamiento de los sistemas políticos. Se trata de un área de conocimiento de reciente aparición, que se mueve entre una versión científico-académica y una versión aplicada cercana a la acción política práctica y pragmática.

Esta actividad se ha abordado desde distintos enfoques, [Fag66, Mea76, NC], que en conjunto aportan un conjunto de características de este área:

- Una definición amplia de comunicación, que incluye la transmisión de símbolos e información, junto a opiniones, comportamientos, hechos y circunstancias de contexto.
- Un concepto amplio de la política entendida como la actividad que regula las relaciones entre las personas y los colectivos.
- Una actividad con multiplicidad de actores.
- Como un proceso de interacción .
- La comunicación política que combina el corto y el medio plazo, esto es, como una actividad estratégica.

La construcción social de historias compartidas y su difusión a través de símbolos y del lenguaje forma parte del ser humano. Los símbolos son los instrumentos básicos para la comunicación. A medida que el ser humano ha ido evolucionando, la comunicación ha ido haciéndose menos mítica, alejándose de lo inalcanzable, y lo mismo ha sucedido con la iconografía política, que se hecho más cercana. Lo político tiene mucho de manejo de lo simbólico, y no solo de lo instrumental. La técnica de contar historias o *storytelling* supone la narración y puesta en escena de relatos colectivos con el fin de mantener con ellos el poder o conquistarlo.

Como sintetiza Canel, la comunicación política adquiere varias dimensiones claramente relevantes [Can99]: (i) una dimensión teleológica o intencional hacia unos resultados concretos y prestablecidos; (ii) una dimensión axiológica, esto es, los actores cuentan con unos valores, principios de acción y normas, y el resultado de la acción comunicativa puede producir cambios o nuevos valores; (iii) una dimensión afectiva, de modo que el cambio en el comporamiento puede

afectar a la dimensión emocional de las personas; (iv) una dimensión rutinaria o procedimental, que implica que la comunicación está incluida y naturalizada dentro de la vida cotidiana de las personas; (v) una dimensión de acción dramatúrgica en tanto que es un acto de representación, que mezcla lo real con lo irreal, e integra reflexiones racionales con decisiones irracionales; y (vi) finalmente una dimensión de acción comunicativa por cuanto que un emisor trata de transmitir y compartir significados con sus destinatarios.

Son tres las principales corrientes de pensamiento en las teorías acerca de la influencia de la Comunicación Política en general y en concreto de las campañas electorales, a saber: las teorías de comunicación de masas, las que pivotan sobre la identificación partidista y por último las relacionadas con una dimensión cognitiva de la comunicación. Mientras las teorías acerca de la propaganda de masas, anteriores a la Segunda Guerra Mundial, se fundamentan en una gran acción propagandística sobre los colectivos entendidos como un sujeto uniforme, las teorías posteriores a ella, están basadas en la identificación partidista de los ciudadanos. Tras estos enfoques, teorías más recientes abogan por la importancia clave de los efectos cognitivos, de agenda-setting y de persuasión de la comunicación política [CM02].

Durante los procesos electorales tienen duraciones limitadas en el tiempo por razones normativas, perso la comunicación y la práctica políticas han derivado en los últimos años en los que se denomina *campaña permanente* [OM00]. Durante los procesos electorales, la comunicación política sigue estrategias diferenciadas a las de la lógica no electoral. Durante la campaña, la publicidad es la mayor fuente de información para los votantes. Durante estos procesos, la comunicación trata de encontrar un equilibrio entre la información nueva (aportada durante la propia campaña) y la información conocida. Comprender las reacciones de los votantes a dichas informaciones es fundamental para entender la eficacia de las campañas. Muchos investigadores que estudian esta eficacia lo hacen confiando en indicadores verbales para medir la influencia de dichos anuncios [Iye10].

En general se ha comprobado que la mayoría de los anuncios políticos polarizan a las personas fieles a los partidos (partisanos o *partisan* en inglés), lo que lo hace consistente con investigaciones previas que muestran que la exposición a la campaña refuerza las predisposiciones de los partisanos que ven los anuncios [AI95]. En el caso de sistema político bipartidista como el estadounidense, los partidarios o partisanos de uno y otro color político lo hacen rápido que los independientes, que en muchas ocasiones apenas se mueven expresando su conformidad o disconformidad [Iye10]. En particular, el cambio se produce más rápido para los demócratas, aunque la velocidad de esta adaptación depende de

la temática de los contenidos.

2.3.1 **De la comunicación de masas a la comunicación interpersonal en redes**

La actual comunicación política forma parte de la denominada tercera era de la comunicación. Emerge con la proliferación de los grandes medios de comunicación y se caracteriza por la abundancia de medios, la ubicuidad, la investigación y la celeridad de los procesos comunicativos sus efectos políticos.

A lo largo de las últimas décadas surgen los servicios de noticias durante las 24 horas, proliferan los ordenadores personales con capacidad de comunicación desde los hogares. Desde el lado de los periodistas, comienzan a acortarse los ciclos de generación de noticias se acortan, con lo que la reflexión y la valoración de la información que se cuenta disminuye.

Durante esta última tercera etapa, se intensifica la profesionalización del apoyo político, incorporando nuevos perfiles profesionales especialistas en marketing, comunicación o periodismo, entre otros. Esto especialmente relevante en Estados Unidos, donde estos equipos y organizaciones se convierten en una verdadera elite política. Durante estos últimos años han aparecido perfiles para la gestión de las comunidades virtuales o *community managers*), que construyen y modulan el mensaje comunicativo. Unido a un proceso de des-mediación en la comunicación política que tiene lugar en la esfera digital [Ras97, Bim98, Ort11].

Este nuevo modelo de comunicación profesionalizado se constituye como un verdadero sistema de comunicación fragmentado en el que se planean con todo detalle desde las campañas electorales, las respuestas a ofrecer diariamente, las tendencias de opinión y se definen los marcos que los periodistas después utilizarán casi de manera necesaria en sus contenidos informativos. Por su parte, en Europa, los estrategas de los partidos tratan de sacar lecciones de las campañas estadounidenses, entre las que destaca que en la práctica totalidad se profesionaliza la publicidad política.

Lograr el éxito electoral requiere, en consecuencia, una elaborada preparación y coordinación, decidir las salidas a los medios y sus formatos, el ciclo de pre y post eventos para asegurarse que la información entendida y transmitida por los periodistas es las que se desea. El objetivo de los departamentos de comunicación de los partidos consiste en llegar a los medios, para que informen de la agenda del partido en cuestión en los términos que a éste le interesan, y que coloquen la crítica en el adversario político. Se adopta el término de *campaña permanente*, que se extiende a todo el tiempo político, sin interrupción.

Todos los eventos se preparan para ser *mediatizados* [BK99], esto es, incluidos en la planificación y agenda de los medios, que los transmitirán finalmente a la opinión pública.

Según Blumler, esto ha traído consigo algunos otros efectos problemáticos [BK99]:

- La subordinación de las políticas públicas a los imperativos de campaña y a la presión de los medios, dando lugar a la denominada *Escuela de la respuesta instantánea*.

- La dudosa relevancia que mucha comunicación política tiene con respecto a las tareas sustantivas de gobierno y los asuntos que les interesan a los ciudadanos.

- Se ha generado cierto rechazo, indiferencia o escepticismo hacia las campañas electorales, dado que mucha comunicación política tiene un marcado carácter negativo, demasiado centrado en luchas internas, demasiado programadas y repetitivas y carentes de credibilidad convincente. Por ello nace la *política de la confianza* que se sustancia en el doble perfil *campaña-sin-candidato* y *candidato con otro perfil* y que, frente a otras opciones, muestra a un líder confiable, con una intensa defensa de los valores y de la integridad.

Por otro lado, con respecto a los aspectos meramente informacionales, el modelo de campaña electoral también ha sido objeto de crítica, desde dos puntos de vista principalmente:

- *Selección a la exposición*. Los ciudadanos no se exponen a todo tipo de información política, sino que seleccionan la fuente de información. De esta manera, si esto es así, los efecto de la campaña electoral serían mínimos en la medida en que sólo servirán para reforzar o movilizar en su caso, el voto, que vendría definido con anterioridad. A veces no es posible una exposición selectiva (por ejemplo, por la sobredifusión que en los medios hace que casi sea imposible no escuchar un mensaje electoral). Según algunos autores, esta imposibilidad de exposición selectiva, aunque no tenga efectos a corto plazo, los tiene a medio y largo plazo.

- *Baja explicitación en muchos mensajes*. En este capítulo cabe citar la acción política del gobierno se traduce en acción comunicativa marketing público e institucional durante la pre-campaña, con mensajes sobre la gestión pública del partido en el gobierno. Este tipo de comunicación ha sido

> regulada en España mediante la Ley 29/2005, de 29 de diciembre, de Publicidad y Comunicación Institucional, que trata de fortalecer la transparencia de las campañas de publicidad y comunicación institucional del sector público y evitar usos indebidos de este tipo de comunicación.

Con el despegue de internet como canal de comunicación, se habilita un nuevo entorno competitivo entre medios y comunicadores políticos en el que los actores del periodismo político ya no dominan un nuevo espacio diferente al que antes operaban. La progresión de *Internet* ha sido veinte veces más rápida que la del teléfono, diez veces más rápida que la radio y tres veces más rápida que la televisión [PP09]. En esta nueva realidad, los sitios web juegan un nuevo espacio en el que los ciudadano toman lo que consideran y cuando ellos deciden, en muchas ocasiones basados en la relevancia de un tema de interés particular o general o la conveniencia, más allá de la lealtad de marca.

Sobre los sistemas de medios de comunicación tradicionales, estos cambios tienen consecuencias de notable envergadura:

1. La cobertura política trata de ser compatible con los propios objetivos de los medios de comunicación.

2. Se reconoce que las aproximaciones más *adoctrinadoras* a la política tienden a no ser válidas para audiencias de masas heterogéneas.

3. Mucha de la competición entre medios proviene del entretenimiento, por lo que comienzan a proliferar aproximaciones a la política basadas en *infotainment* o *infoentretenimiento*. Se produce una explosión de subgéneros y formatos híbridos: programas de desayuno o *breakfast shows*, espectáculos hablados o *talk shows*, revista de noticias, investigaciones de escándalos políticos, televisión *tabloide*, entre otros.

4. Las formas tradicionales de periodismo político sufren gran presión, incertidumbre y erosión. La prensa, nunca homogénea, se fragmenta con actores más pequeños en *Internet* y televisión. La ruptura de la división público-privado divide la cobertura a los políticos, con un significativo ascenso de los escándalos como tema y como atractor de audiencia.

Estos cambios suceden con mayor intensidad en Estados Unidos, mientras que en Europa no se produce en el mismo grado, debido a dos factores principalmente. Por un lado, culturalmente, el público europeo aún se mantiene más receptivo a las agendas de los políticos serios. Por otro lado, los sistemas de medios y la competencia multicanal es muy diferente, contando en Europa con organizaciones y servicios públicos de comunicación ya consolidados.

En general, se usa el término *americanización* [SM96] de las campañas electorales y de la política, desde la década de los años 60 del siglo XX, cuando la televisión condujo un modelo de campaña electoral con alcance nacional, junto con una profesionalización de los equipos de campaña, otorgando una importancia destacada al marketing de la imagen [FKM01].

2.3.1.1 **Campañas en *Internet***

Desde la aparición de *Internet* y su popularización para uso cotidiano, actores políticos como las instituciones o los partidos han incorporado este medio como uno de sus instrumentos de comunicación con su electorado y con la sociedad en general. Supone una enorme red de redes configurada como un espacio abierto que [Cas09] provee un arquitectura para la participación. Se trata de una estructura sin jerarquía cerrada para lo político, y que permite una comunicación tanto de arriba a abajo, como desde abajo hacia arriba [O'R07].

La primera experiencia reseñable del uso electoral de *Internet* tuvo lugar en Estados Unidos durante la campaña a las elecciones como gobernador de Minnesota en 1998, a cargo del candidato Jesse Ventura. Como candidato hizo uso de esta red de redes para publicitar los mítines de cierre de campaña. Posteriormente, el senador John McCain utilizaría la misma red como vía para recaudar más de tres millones de dólares [Klo04]. Sin embargo, fue la campaña presidencial de Howard Dean en 2004 la que introdujo cambios significativos y sentó las bases para gran parte de las estrategias digitales de las campañas electorales a partir de ese año. De ese modo se convirtió en el candidato que más fondos recaudó en la historia del partido demócrata [Hin10].

La utilización electoral de las redes sociales fue la gran protagonista durante las elecciones de media legislatura (*midterm elections*) en Estados Unidos en 2006. En ellas, *YouTube* se convirtió en plataforma electoral para la difusión de vídeos con declaraciones de candidatos que comenzaron a propagarse de manera viral. Posteriormente, las primarias de 2007 y 2008 fueron escenario del auge de los sitios de interacción social en red como *MySpace*[1] y *Facebook*, donde cientos de miles de simpatizantes y partidarios tejieron una densa red de apoyo a los candidatos demócratas.

Internet ha servido para crear redes políticas, promover la discusión sobre asuntos políticos, recaudar fondos para las campañas, así como para automatizar las tareas de obtención, almacenamiento, recuperación y envío de información a todas las personas que participan de estas redes [How06].

[1] `http://www.myspace.com`

La relevancia de los medios sociales online en la comunicación electoral en distintas geografías he quedado evidencida, como en el caso de Estados Unidos [GGR10, Smi09a] o Europa [Jac07b, JL11, Kal09, LJ10].

Desde una perspectiva funcional, se ha estudiado cómo y por qué los partidos políticos eligen un medio particular para un objetivo concreto [Te'01, Web98]. Así, desde la "teoría de la riqueza de los medios" [DL86] se explica el uso de *Internet* como elemento estratégico en las campañas electorales. Según esta teoría, cuando el mensaje es complejo existen múltiples obstáculos para que los receptores lo entiendan, por lo que el medio a través del cual se transmite ese mensaje debe ser "enriquecido". Si la conversación cara a cara es el canal más "rico", la prensa escrita es no muy rica, mientras que *Internet* se sitúa en la zona media [Te'01].

El proceso electoral de las presidenciales estadounidenses de 2008 fue la primera vez que los partidos trataron de definir sus campañas como *conversaciones* con los ciudadanos americanos. Esta conversaciones formaron parte de la estrategia de los partidos, lo que supuso un cambio con respecto a las campañas tradicionales. En ellas, el enfoque *multicanal* abarca desde los anuncios de televisión, las caravanas de campaña, el correo postal y la cobertura en prensa, pero dejaban de lado la estrategia digital.

Si el propósito en todo caso es controlar el mensaje y el tono a través de métodos unidireccionales de comunicación [Wes07a], la única forma de obtener información retroalimentada (del inglés *feedback*) era mediante la técnica de cuestionario. Sin embargo, desde la campaña de Howard Dean, los estrategas se dieron cuenta de que se podía medir el interés y la opinión a través de Internet, además de propiciar la compromiso político [Hin07]. Esta es, en parte, la filosofía de la Web 2.0. Con plataformas como *Facebook* es posible organizar y comunicarse son los seguidores de un modo muy eficiente. Asimismo, proporciona miembros del ámbito público la capacidad de difundir sus opiniones y organizarse de manera independiente.

En España, se ha estudiado cómo los partidos políticos han hecho uso de las tecnologías basadas en *Internet*, tanto durante las elecciones generales como las autonómicas en Caytaluña [LG11, VO12, DCC$^+$14, DB14]. A partir de estos trabajos se pueden sintetizar los principales instrumentos tecnológicos utilizados:

- **Blogs**. Durante las elecciones generales de 2008 irrumpe el fenómeno blog, con la proliferación de multitud de espacios de opinión personal y análisis político por parte de partidos políticos, periodistas y también de ciudadanos. Los blogs son *webs* que publican contenidos a título personal (por ejemplo, de un candidato) y que permiten cierto grado de interac-

ción con su audiencia. La constelación de blogs sobre asuntos políticos constituye la denominada *blogosfera*[1] política [AG05]. Los partidos lo incorporan dentro de su estrategia de campaña. Se consolidaron los blogs como medio de información y de propaganda política, al mismo tiempo que tiene lugar un gran desarrollo de las webs de los políticos.

Los medios de comunicación en su faceta *online* concentraron a la mayoría de internautas, por encima de las webs de los partidos y los blogs [Ric08]. Los grandes actores en la *blogosfera* durante dichas elecciones fueron, por un lado, medios de comunicación que no eran nativos digitales, y por el otro, redes de ciudadanos que orbitaban alrededor de una red ideológica [LGVG11].

- **Redes sociales**. En 2008 los grandes partidos comienzan a tener presencia en las redes sociales, como Facebook[2] o *Twitter*[3], donde no sólo los partidos, sino los candidatos empiezan a desplegar su repertorio de acción comunicativa.

  Pero el mayor despliegue de una estrategia digital hasta la fecha es "Política 2.0" a manos del PSOE, que a través de comunidades de blogs (que suman más de 800 blogs) y otras herramientas de unificación y amplificación de su voz concentra y conecta a miles de usuarios a través de la web. A través de esta *blogosfera* progresista se pretendía llegar a los simpatizantes y militantes a través de una estrategia de marketing viral dirigida a todos los ciudadanos.

- **Vídeos**. Especialmente a través de la red social *YouTube* fue una de las armas clave en las campaña de 2008 y 2011. En la campaña de 2008 aparece un nuevo género, la *videopregunta* al candidato, convertida en una suerte de "videocracia", a la que se sumaron medios de comunicación como Antena 3 o La Sexta. Recogía en cierta manera el formato inaugurado con el programa "Tengo una pregunta para usted" del canal de televisión La 1.

En consecuencia, los principales rasgos de esta práctica de comunicación política basado en tecnología se pueden resumir de la siguiente manera:

[1] Blogosfera es la denominación genérica de cualquier red de *blogs* sobre un tema o área que están interconectados mediante enlaces.

[2] El PP a través de su cuenta `https://www.facebook.com/pp` , y el PSOE a través de su cuenta `https://es-es.facebook.com/psoe`.

[3] El PP a través de su cuenta @PPopular, en `https://twitter.com/ppopular` , y el PSOE a través de su cuenta @PSOE en `https://twitter.com/psoe`.

- **Uso estratégico de la tecnología**. Las herramientas tecnológicas han sido utilizadas de manera distinta a como se habían utilizado antes (ligadas a la expresión y movilización de los movimientos sociales), sino como auténticos instrumentos de marketing, principalmente a través de los canales que los dos grandes partidos habían creado en sitios sociales como *YouTube*, *Facebook* e incluso *SecondLife*[1] [LGVG11, SP08].

- **Replicación de modelos**. Desde las elecciones de 2004 y con mayor intensidad en las generales de 2008 y 2011, los partidos políticos en España han venido teniendo una presencia digital que es producto de dos elementos. En primer lugar, los usos son prácticamente idénticos a los de otros partidos políticos en otros países occidentales, y por otro lado han mantenido los criterios en su acción política digital, incluyendo lo ideológico, lo mediático e incluso la relación con los propios ciudadanos.

- **Desintermediación**. La principal aportación del proceso de comunicación a través de redes sociales radica en que, además de otros propósitos, se está propiciando una ruptura del monopolio de la intermediación que los grandes medios de comunicación habían tenido durante muchos años, fragmentando dicha mediación y multiplicando el número de transmisores, y como consecuencia de lo cual resultaría mucho más difícil de controlar.

- **Discusión política basada en el alineamiento partidista**. La armazón política estriba en que ambos grupos plantean líneas ideológicas enfrentadas, dejando al margen la discrepancia y por tanto, minimizando cualquier resquicio de diálogo y confrontación dialéctica. Confirma la tesis de Jenkins acerca del papel de los blogs como instrumentos de reafirmación y cohesión ideológicas orientado a la movilización [Jen08]. Aproximadamente una cuarta parte de los ciudadanos siguieron a diario la campaña a través de *Internet*, mediante distintos tipos de actividades, tales como la consulta de blogs y visualización de vídeos, realizando encuestas *online*, y por medio de la lectura y publicación de comentarios en distintos foros. *Internet* entra en la campaña, no sólo como medio en el que se desenvuelve la entrega de información, la propagación ideológica, sino que además lo hace como temática de debate y confrontación. El nivel de discusión ciudadana dista bastante de ser considerado democrático. Aun así, existió una gran diverisdad, una baja reciprocidad o nivel de respuesta a los mensajes expuestos [ODG13].

[1] `http://www.secondlife.com`

A lo largo de la campaña de 2011, el tamaño de la *blogosfera* política se mantiene, si bien es cierto que el activismo en las redes sociales cobra un especial interés. En ambas elecciones, las opiniones y el debate que ocurre en los blogs está polarizado, confirmando las algunas de las tesis pesimistas sobre la deliberación en *Internet* [AG05] y sobre la fragmentación y polarización ideológicas [Sus07].

- **Correspondencia con el sistema de medios**. Durante campaña de 2011 también se pone de manifiesto de nuevo la existencia de vínculos entre actores políticos y medios de comunicación, expresado a través de elevados niveles de paralelismo político entre sus respectivos mensajes. El sistema político-mediático español se caracteriza por un alto grado de pluralismo polarizado, con una fuerte asociación entre partidos políticos, instituciones y grupos mediáticos. Esto ha conllevado tradicionalmente a audiencias que generalmente son partidistas, lo que supone un auténtico factor clave a la hora de explicar el consumo de medios [HMW08].

- **Estrategia vertical *top-down***. El espacio de una denominada nueva *política 2.0* en la que Internet actúa como espacio para la reflexión da paso al uso partidista de estos canales, aumentado por la reproducción de sus mensajes a través de los canales digitales de los medios de comunicación. Este esquema supone repetir el modelo jerárquico de transmisión de información tradicional de arriba a abajo.

Durante las elecciones generales de 2011, los movimientos alrededor del 15M supusieron un cambio con respecto a los procesos anteriores. Estos movimientos confluyeron en la aparición de un nuevo actor político, el partido Podemos. A partir de un elevado grado de tecnificación o tecnologización de la política, que ha supuesto un cierto empoderamiento de los ciudadanos, no a través de los partidos convencionales del sistema de partidos, sino a través de otros vectores de expresión como el partido citado.

Recordando la tipología de perfiles en Internet de Sampedro [Sam11], hemos de destacar describe tres tipos de perfiles en función del tipo de actividad que llevan a cabo en la red. En primer lugar, los *ciberactivistas sin fundamento* reenvían información sin analizar o reflexionar sobre su contenido (a veces sin consultarlo) o sus efectos. Este tipo de comportamiento supone un ejercicio de banalización de la ciberpolítica.

En segundo lugar, los *ciberactivistas lúdicos* generan y reenvían gran cantidad de contenido de *infosátira* (esto es, que transmite información desde la burla o sátira), junto a parodias de los candidatos y de los partidos, sobre propuestas partidarias a destinatarios con cierta afinidad ideológica.

En tercer lugar, los ciberactivistas molestones o *trolls* publican y transmiten contenidos lúdicos o con distinto grado de provocación o molestia. Es la figura del *trol*, ya estudiado por [TSFW05, Har10, DlP14]. Este rol, desarrollado en el ámbito de Internet, pretende introducirse en conversaciones de terceros con el propósito de interrumpir el fluir natural del grupo. Por último, existe otro rol desempeñado por personas que solo quieren manifestar su opinión en un blog o en un foro. Principalmente son jóvenes y no existen evidencias claras de que tengan una gran componente de compromiso cívico y/o político. Este es el rol de los ciberactivistas expresivos.

En cuanto a la pertenencia a algún tipo de organización, se distinguen distintos perfiles de ciberactivistas:

- Los *ciberactivistas implicados* promueven su afinidad a un partido o ideología hacia otras personas. Aunque no forman parte de la maquinaria de ningún partido, propagan esta afinidad reenviando información o promoviendo la movilización política (tal como asistir a una manifestación o mítin) y también elaborando material audiovisual a distribuir en sus mensajes.

- Los *ciberactivistas de partido* forman parte de las estructuras de los partidos, y dedican parte de su jornada a elaborar contenidos, publicar mensajes y reenviar esta información a través de canales como el correo electrónico, los blogs, los foros o las redes sociales. Al servicio de sus organizaciones, trabajan sobre con una línea editorial en forma de argumentario.

- Los *ciberactivistas sociales* están vinculados a movimientos ciudadanos y movimientos sociales, por lo que tratan sobre distintos temas, manteniendo posiciones distintas sobre temas de sociedad y política. Aunque tienen no pertenecen estrictamente a la lógica de partidos, sí tienen relación con las posiciones ideológicas de éstos, por lo que han de considerarse.

Todos los anteriores tipos de ciberactivistas comparten una visión similar respecto de *Internet*, ya que les proporciona un medio con mayor autonomía, un grado menor de estructuración -y por tanto menos restrictivo-, que permite niveles mayores de interacción y que hace posible la proactividad de los participantes. En el mismo sentido, se entiende que *Internet* ofrece un modelo de comunicación y acción política más dinámico, frente a la acción política más convencional, caracterizada por plantearse de arriba abajo, orientado al marketing político. Se entiende Internet, no como un mundo aparte, sino como una parte más del mundo.

En consecuencia, en España se constatan algunas adversidades al modelo de ciberpolítica [Sam11]:

- El carácter *virtual* de la red, que se manifiesta a través de un impacto reducido, con una penetración social limitada, principalmente por el sesgo y la barrera generacional, además de la pobre cultura participativa que existe en la ciudadanía.

- La ausencia de un discurso propio sobre la ciberpolítica, más allá de la función meramente simbólica de las TICs, las consecuencias limitadas que se pueden generar o la creciente espectacularización y negativismo de la ciberpolítica, que de este modo se alinea con el resto de campañas en las elecciones generales.

- Los discursos más elaborados y profundos se dan a medida que crece el grado de ciberactivismo y la identificación ideológica, frente a las conversaciones más triviales que tienen lugar cuando la interacción es corta en intensidad, de duración breve o con un número reducido de participantes.

- A nivel deliberativo, la crítica más sofisticada a los argumentos de los demás usualmente está en relación directa, principalmente, al grado de conocimiento político, y además, al uso de las herramientas y tecnologías que permiten dicha discusión o debate político.

2.3.1.2 **Deliberando a través de una comunicación bidireccional**

Combinando el enfoque instrumental que se acaba de describir -las elecciones- con una mirada de mayor alcance -la democracia-, el concepto e implicaciones del *sitio social digital* tiene el potencial para expandir el concepto de "esfera pública" elaborado por Habermas [HLL74] en la medida en que permite a lo público comprometerse en la acción política tanto en su globalidad como de manera independiente y particular durante las campañas políticas.

Este concepto de *esfera pública* supone como un buen punto de partida: un lugar donde los miembros de una comunidad podrían colectivamente formar una opinión pública en un entorno en el que no existe ni gobierno ni economía. Habermas no explica cómo los políticos y las organizaciones podrían encajar en la mezcla. Pero, como el propio Habermas afirma, para una comunidad que realmente fomente la comunicación política, deberá ser mucho más inclusiva.

Una comunidad con buena comunicación política requiere participación de los actores políticos, los medios y el público; los actores políticos hacen llegar

su mensaje bien directa o de manera mediada hacia la comunidad. Los medios deberían traer información al público, además de comprobar al gobierno, investigando y validando la información que estos les proporcionan. Los miembros de la comunidad deberían tomar un papel activo, además de cuestionar las fuentes, responder a los periodistas y trasladando las historias verdaderamente relevantes.

Desde la democracia deliberativa del propio Habermas, el acceso a información adicional debería ser fácil y sencillo. Los representantes no sólo deberían dar cuenta en las elecciones, sino en cualquier sitio y momento, y los miembros podrían requerirles explicaciones en esas mismas condiciones. En consecuencia, el umbral para la participación política debería ser bajo, permitiendo que los planteamientos y propuestas de cualquier ciudadano puedan ser escuchados, pudiendo llegar a convertirse en líder de opinión. En ese mismo escenario, sería posible, asimismo, que los ciudadanos puedan organizarse, planificar eventos, crear grupos para dar mayor peso a sus opiniones. De esa manera se estarían facilitando que se crearan movimientos de base a partir de una idea [Hab89].

La comunicación tradicional de los partidos ha sido unidireccional: surge desde la organización del partido y se dirige unidireccionalmente hacia sus destinatarios, bien sean simpatizantes o electores. Estos pueden ser su propia organización (en las figuras de afiliados o simpatizantes) o bien los ciudadanos. [WGC+98, SR03, YP98]. Sin embargo, durante los últimos años se ha producido una transformación en el modo en que esta comunicación opera entre emisores y destinatarios, de modo que los partidos han tratado de asegurar que sus votantes escucharan los mensajes que se quería que conocieran [Col01] y [JPW01]. Sin embargo, el mero hecho de que exista bidireccionalidad en la interacción no implica necesariamente una conversación efectiva y real. De este modo, la asimetría en la bidireccionalidad hace que mensajes escuchados no sirvan para que el destinatario modifique o module su respuesta, por lo que niveles mayores de simetría han mostrado niveles mayores de conocimiento y comprensión de los mensajes [GH+84].

2.3.2 **Facebook como canal de comunicación política**

Lanzado en 2004 como aplicación web, *Facebook* se ha convertido en la comunidad *online* que conecta más miembros del mundo real de manera más efectiva. Los políticos han venido usando *Facebook* para comunicarse con los miembros de su comunidad que desean escucharles. En cierto modo, les agrupa y les proporciona un medio para compartir información a través de una red accesible desde multitud de dispositivos, conectándose a sus áreas de interés y organi-

zándose en torno a ellas. Por el otro lado, permite a los actores políticos un medio eficaz de llegar a los electores y votantes, aunque no puede conseguir que la información de éstos sea procesada y produzca un efecto en los políticos.

Uno de los ejemplos más relevantes es el de las campañas de primarias y presidencial de Barack Obama en la carrera hacia el gobierno de Estados Unidos [Nor00], y los efectos de su estrategia digital a través de medios sociales *online*.

En las siguientes líneas se describen algunos de los principales rasgos del uso partidista de *Facebook* en las campañas electorales.

- **Comunicación política entre los miembros de *Facebook***. La herramienta más comúnmente utilizada para la comunicación política son los "grupos" *Facebook*, en los que pueden recibir información, invitar a otros a unirse, publicar o *postear* su propia información, reforzándola o enriqueciéndola con fotos, vídeos y enlaces a otras informaciones y conversaciones.

  También existen grupos partidistas, en los que la conversación no es tan abierta, sino que se estrechan los temas y se ignoran puntos de vista divergentes o diferentes. Apenas existen grupos en los que conviven información y temas sobre los que existen opiniones claramente controvertidas. En este sentido, *Facebook* actúa como una infraestructura de comunicación, pero no necesariamente la estructura de una comunidad de un partido [1].

  Esta red permite compartir una historia de noticias, vídeos o una página web con los demás. Los miembros de un grupo pueden incluir comentarios, compartirlos con terceros (sus amigos) o proporcionar fuentes de información complementaria. De este modo, los partidos políticos pueden lleva a cabo tareas de información, divulgación y transmisión de posicionamientos de su partido entre sus seguidores y, a su vez, éstos difundirlos a terceras personas más allá del grupo. Así mismo, estos mensajes cruzados pueden suponer la puesta en común de opiniones distintas, argumentaciones o simplemente la exposición de información.

- **Comunicación política entre miembros *Facebook* y políticos del mundo real**. Los candidatos incluyen sus biografías extendidas y resúmenes en sus páginas *Facebook*, información de contacto. Su repertorio de acciones incluyen el uso de *posts* tipo blog y reciben *feedback* de sus seguidores políticos. Para la mayoría de los miembros, muestran en *Facebook* su actividad personal, como Mitt Romney en 2008, cuando dedicó mucho tiempo a su

[1] *Facebook* se comercializa a sí mismo como un concentrador (*hub*) de interacción social.

perfil, con muchos álbumes de fotos, eventos y enlaces a artículos en los periódicos.

La mayoría de perfiles dejan de estar activos después de la campaña y otros no mantienen contacto con los electores después de la misma, cuando la mayoría prefieren usar el sitio web oficial para dicha comunicación. Adicionalmente, *Facebook* permite a los miembros de una comunidad poder enviar mensajes directamente a los candidatos, aumentando de tal modo el nivel de interacción bidireccional.

- **Facebook como herramienta de movilización**. Uno de los principales beneficios de *Facebook* ha sido ser un medio para movilizar y organizar miles de apoyos. Si Howard Dean consiguió, a través de MeetUp.com[1] 190.000 miembros en las primarias de 2004, estos grupos tenían el objeto de organizar reuniones del mundo real para tratar/discutir sobre un tema mientras que los grupos de FB no tienen esta componente de movilización.

  Existen muchas formas en que los miembros puedan utilizar las funciones de *Facebook* para organizar eventos políticos: los administradores de los grupos distribuyen/lanzan mensajes a los miembros para recordarles las novedades de la campaña, creando de facto una lista de seguidores voluntarios; los miembros invitan a otros a unirse o remitir mensajes en las campañas. Esta plataforma es una valiosa herramienta para informar, movilizar y organizar a los seguidores políticos; el único aspecto que queda pendiente en la recaudación de fondos.

- **El futuro de la Política Facebook**. Para Gulati "los beneficios reales de los sitios de interacción online se notará en el futuro para aquellos candidatos que continúen comprometiendo a sus seguidores después de terminar las elecciones promocionando una sensación genuina de comunidad entre ellos". En la misma línea reconoce que casi ninguna de las características de Facebook es completamente original, pero es su combinación lo que le hace efectivo a *Facebook* [WG09].

  Los estudiantes universitarios, a menos que sean abiertamente políticos, no visitan sitios web políticos, por lo que se llega a ellos a través de los campus en *Facebook* o *YouTube*. Por tanto, no se trata tanto de atraerlos a la política –a los partidos- como el de acercar/integrar la política en su interacción social *online*.

  Proporcionando un espacio para que los ciudadanos interpreten la información política y participen en debates políticas, esta red social se ha

[1] Meetup.com es una red social digital creada en 2002. `www.meetup.com`

> convertido en un gran añadido al mundo de la comunicación política. *Facebook* no se va a convertir necesariamente en el primer medio para la comunicación política, pero su valor estriba en el poder de los miembros para organizarse y responder ante una situación pública.

En España, el marco de trabajo de Di Bonito analiza ofrece un doble enfoque [DB14]. En primera instancia, desde la perspectiva funcional, muestra una utilización de Facebook por parte de los partidos orientada primordialmente a proporcionar información política, con pocas apelaciones directas al voto o a la movilización. En segunda instancia, desde el análisis argumentativo, se confirma el grado de personalización de los contenidos hacia los propios candidatos y la autorreferencia de los contenidos elaborados por los propios partidos. Las conversaciones se inician desde los muros *Facebook* de los propios partidos, no permitiéndose que sean los ciudadanos quienes abran la discusión (con la excepción de *Ciutadans*[1]). Sin embargo, el nivel de respuesta es muy reducido, por lo que no se puede afirmar que supongan un vehículo para canalizar y abrir conversaciones o debates en profundidad .

Se señala la influencia de los medios sociales sobre el apoyo a los candidatos y los partidos, así como su contribución facilitando la participación electoral. La mayoría de los ciudadanos no se preocupan de acceder a los foros en los que los partidos informan de sus propuestas políticas para el gobierno, y cuando lo hacen ven cómo los partidos ofrecen información muy vaga y pobre [AI95, PM76, Wes97].

2.3.2.1 **Los mensajes en las redes**

El análisis de mensajes en redes sociales ha sido objeto de estudio para conocer el proceso de reclutamiento de un movimiento social dentro de redes sociales online [LGA+11, GBBHRM11], la movilización política [BFJ+12] o cómo fluye la información de los partidos en redes online y la formación de comunidades de ideología política [CGFM12, WPF+09, AG05, CFME11]. Así, algunos autores han utilizado el análisis de los flujos de información como termómetro en tiempo real o para anticipar o predecir resultados electorales [TSSW10]. Otros trabajos han planteado también si la propia estructura de la red, esto es, la mera arquitectura de las conexiones entre sus nodos, puede condicionar la influencia política o la formación de grupos de personas afines a un partido político [CGFM12, CFME11].

[1] *Ciutadans* es la denominación inicial del partido Ciudadanos. Durante su etapa inicial, su ámbito territorial de acción política se limitaba a Cataluña.

Aún así, pese a que los estudios sobre redes sociales así lo consideran, no existe un consenso completo sobre si un *Like* en *Facebook* o un *Retuit (RT)* en *Twitter* implican un respaldo o reafirmación al mensaje [BGL10].

El flujo de información comporta un alto componente partidista en tanto que las conversaciones se organizan en torno a los perfiles de los partidos políticos y de los candidatos, en línea con lo que sucede en Estados Unidos [CGFM12]. Este sesgo conversacional solo se rompe ligeramente cuando las conversaciones se establecen con perfiles de partidos con los que se comparte cierta identificación ideológica. No se encontró una correlación relevante entre el volumen de las conversaciones en torno a determinadas comunidades con respecto a los resultados electorales del partido vinculado a dichas comunidades [dEPyC].

2.3.2.2 **Exponiendo las diferencias políticas**

Uno de los principales papeles de los sitios de interacción social ha sido proporcionar espacio para expresar opiniones y participar en algunas actividades (*The Pew reseracrh Center*), y en especial propiciar un carácter más activo de los ciudadanos durante las campañas [Smi09b]. Diversos trabajos han analizar las distintas influencias de los sitios de interacción social, tales como la gestión de impresión que los contenidos publicados generan en la audiencia, la estructura de la red de nodos conectados, o aspectos de privacidad [BE07].

Los principales trabajos se centraron en la conexión entre política y sitios de interacción social, y más específicamente en actividades de participación como el *civic engagement* o *compromiso cívico* y el capital social [ESL07][PKV09]. Aún así, hoy en día se mantienen diferentes posturas sobre si estos sitios web contribuyen a la democracia pluralista. En otras palabras, si las los sitios de interacción social favorecen la expresión de distintos puntos de vista políticos transversales [Mut02a, Mut02b, WM09] o entre contactos heterogéneos [Bru10, SNBN04].

Más allá de lo que sucede en la arena digital, los estudios electorales clásicos ponen de manifiesto que una relación entre la exposición ante puntos de vista diferentes fomenta actitudes más fuertes y mejora la capacidad de movilización [Sch82, Str06]. Por el contrario, escuchar únicamente posturas similares a las de uno mismo dificulta el análisis crítico de diferentes circunstancias políticas y genera democracias fragmentadas y polarizadas [Sus07].

La tesis de Sustein es que la completa libertad en *Internet* conduce a que los ciudadanos terminen por escuchar y leer contenidos afines a sus propias posturas políticas, "filtrando" solo aquello con lo que más de acuerdo están, lo que en efecto conduce a una cierta fragmentación la opinión pública y la polarización de la política. Esto contrasta con lo que se desde la psicología se

dice. La consistencia cognitiva hace referencia a que las personas tendemos a buscar enfoques que encajen con nuestra propia perspectiva, reduciendo las disonancias cognitivas y haciendo más eficiente el procesamiento de la información [Fes62, SFN08, JS07]. Este progresivo filtrado en la selección progresiva de fuentes y contenidos de información es a lo que Sustein denomina "comunicación aislada" [Sus07]. Iyengar y Hahn lo evidenciaron mostrando la fuerte correlación entre los canales de televisión y la preferencia partidista en Estados Unidos.

La postura de Bimber, menos pesimista, plantea que *Internet* no es un medio tradicional de comunicación, sino que es una plataforma que ofrece múltiples y variadas posibilidades para a la comunicación pueda ser procesada por personas y también por grupos [Bim08]. Abre, por tanto, la posibilidad tanto a una esfera pública digital como a una fragmentación política.

Para Brundidge, las noticias *online* y la discusión política online contribuyen a expandir la heterogeneidad de las redes de discusión políticas [Bru10]. Es más, su trabajo muestra cómo los ciudadanos, pese a no buscar explícitamente información con diferencias políticas, es muy probable que accedan a fuentes de información diversas, lo que en cierta manera contradice las tesis sobre la homogeneidad. Wojcieszak y Mutz [WM09]) llegaron a conclusiones similares, pero referidas a "espacios apolíticos" en los que las visiones transversales se exponen con mayor facilidad, frente a espacios de intercambio de mensajes breves, o *chat-rooms*, y *muros* de mensajes, en los que esta pluralidad está muy limitada.

Los sitios web sociales influencian positivamente en la exposición política a lo diferente. Frente a lo que podría pensarse, sin embargo, las variables sociodemográficas (edad, género, etnicidad o nivel de ingresos) no son predictores de niveles mayores más altos de exposición.

Es la pertenencia o identificación partidista (*partisanship*) el factor que más negativamente influye sobre la exposición a la diferencia política. En la misma dirección, aquellos ciudadanos que usan los medios de comunicación vinculados ideológicamente a los partidos tienden a exponerse a puntos de vista similares a los suyos propios. Esta relación entre medios de comunicación y partidos es especialmente relevante en el caso español, tal como se ha descrito en las secciones anteriores. Por contra, los ciudadanos que no tienen un grado elevado de identificación partidista tienden a mostrar su discrepancia o divergencia en temas políticos con un mayor número de mensajes, mientras que los partidistas lo hacen en mucho menor grado [Kim11].

2.3.3 La comunicación política desde el análisis de marcos

La comunicación es un elemento clave en los procesos electorales. Desde la Teoría de la Elección Racional se otorga una importancia a la influencia de la comunicación durante las campañas electorales. En especial, señala que la eficacia de la comunicación tiene su efecto en la transmisión de esa información, así como la interiorización y construcción de una opinión política, y su orientación final hacia la acción.

Esto es especialmente importante si lo combinamos con el modelo del voto temático (*issue voting*). En la medida en que unos futuros votantes tienen construida una opinión respecto a una cuestión de su interés, la comunicación puede jugar un papel clave haciendo que ese asunto sea de notoriedad suficiente (que llegue a ser un tema conocido) y de interés (que el ciudadano esté predispuesto, sensibilizado sobre ello), que la opinión de los ciudadanos sea una determinada (construyendo socialmente opiniones) y haciendo que esta opinión coincida con la de una de las opciones políticas (por tanto, haciendo llegar con eficacia la opinión de un partido político).

Dada la importancia de la comunicación desempeña en la acción política, ha recibido atención desde distintas aproximaciones sobre su influencia. El análisis de marcos ofrece una perspectiva para conocer los procesos y las estrategias de comunicación desde un nivel meta-comunicación.

Debido al peso de los medios de comunicación sobre la transmisión de información política, en ocasiones, el análisis de marcos ha confluido con otros enfoques, como el *agenda setting* o definición de agenda. Sin embargo, mientras el *agenda setting* pretende señalar sobre qué temas pensar, el enmarcado o *framing* trata de organizar cómo el destinatario de la comunicación ha de pensar sobre dichas determinadas cuestiones [BD11, Mil04].

Existen fundamentalmente dos corrientes relacionadas con la investigación sobre los enfoques informativos [Neu92, VSDV99]. Un enfoque genérico de los encuadres y un enfoque temático concreto. Los marcos genéricos son aquellos que transcienden la limitación temática y se pueden utilizar para analizar diferentes hechos, a lo largo del tiempo y en distintos contextos culturales. En cambio, los marcos temáticos o específicos (del inglés *issue-specific frames*) hacen referencia a hechos concretos con categorías elaboradas de forma específica. Según ambas perspectivas, un marco supone un principio de organización que transforma la información fragmentaria o casual en un problema político estructurado y significativo, en el que se incluye, implícita o explícitamente, una solución [Ver05].

Los marcos o *frames* son dispositivos que sirven para delimitar la construc-

ción mental y social de un determinado tema o asunto. En el caso de la comunicación política, el proceso de *framing* o enmarcado está por tanto, dirigido a que las personas entiendan y evalúen los acontecimientos y las condiciones en que se producen [IK10] en términos de los marcos o *frames* que se elaboran estratégicamente. Los actores políticos tratan de movilizar a los votantes para que les apoyen (bien sus políticas o las propuestas durante los procesos de campaña electoral), para lo cual incitan a pensar en los términos que consideran más eficaces para que aquellos respalden a éstas. Con este propósito se reproducen y repiten aquellos aspectos que destacan y que están alineados con los valores [Jac00], y lo hacen tratando de construir preferencias de unos temas sobre otros.

El proceso de *framing* se ha abordado desde distintas perspectivas tanto teóricas como metodológicas, lo que ha conducido a una comprensión fragmentada. A finales de los años ochenta y principios de los noventa es la época de la consolidación del estudio teórico de los marcos, además de un reforzamiento de su faceta metodológica. Sádaba *et al.* realizan una clasificación de siete aproximaciones que realizan los estudios de frames: constructivista, estratégica, de actitudes profesionales, metodológica, cultural, narrativa y simbólica [Sá01].

Para resultar eficaces, los frames contienen necesariamente algunos de los siguientes tipos de resonancias [SRVLP08]:(i) resonancias culturales, según las cuales un *frame* es más efectivo en la medida en la que apela a valores compartidos, lugares comunes para una sociedad; (ii) resonancias mediáticas o aquellas que se refieren al modo de trabajo de los medios, esto es, un frame que cumple con ciertos criterios de *noticiabilidad*; y por último (iv) resonancias personales, por las que resulta eficaz incluir los *frames* en la experiencia personal y colectiva de distintos acontecimientos.

2.3.3.1 **El enfoque del encuadre de medios**

El uso que los medios de comunicación llevan a cabo de los marcos ha recibido la atención ha sido objetivo de estudio bajo la perspectiva del análisis de marcen las últimas décadas, y desde distintas perspectivas, tales como el análisis comparado de su uso en medios [SV00], el propósito de interacción de los marcos [TJP$^+$00], o el uso de cara a determinadas políticas públicas [Bar05, HP05]. Se pueden identificar dos niveles de influencia derivados de las características de dos medios como la prensa y la televisión: *slightly permanent agenda* y *less permanent agenda* [BF76].

El enmarcado ofrece una alternativa al viejo paradigma de *objetividad frente a sesgo*, tal y como plantea Hackett, quien propone un terreno para el estudio

de la ideología de las noticias, dejando de lado su enfoque del estudio de la objetividad y el sesgo de la información periodística [Tan01]. De este modo, el concepto de ideología trasciende al concepto de sesgo, de manera que la ideología proporciona un marco de trabajo a través del cual los medios a menudo presentan los sucesos [Hac84]. Los marcos se distinguen del sesgo en los términos siguientes:

- Ofrece un concepto más sofisticado sobre la significación. Va más allá de estar a favor o en contra una determinada afirmación. Añade la posibilidad de respuestas emocionales complejas y proporciona también la dimensión cognitiva (creencias sobre objetos además de actitudes).

- El enmarcado reconoce la capacidad de un texto (o una presentación mediática) para definir una situación, definir los problemas y plantear los términos de un debate.

- Refleja la riqueza de los medios y las sutiles diferencias que son posibles cuando un determinado topic se presenta de distintas maneras.

- Ofrece un medio-herramienta para que los investigadores cuantitativos puedan validar la hipótesis de *hegemonía de medios*, lo que resulta complejo de validar empíricamente. El frame dominante es aquel que la gente acepta sin darse cuenta y/o sin cuestionárselo.

El proceso de encuadre por parte de los medios de comunicación es importante en tanto que puede tener sutiles pero poderosos efectos en la audiencia. En gran medida, esto sucede porque puede definir los términos del debate sin que la audiencia perciba que esto está sucediendo. Y es que la definición del marco es una de las decisiones más importantes que toma un periodista [Smi97]. Por tanto tiene un carácter estratégico.

Desde la perspectiva de la contienda política, según Tarrow, además de la desigualdad económica y las estructuras de oportunidad política existen otras dimensiones de carácter subjetivo, como son las percepciones individuales y la construcción social de significados [Tar97]. Sobre este segundo tipo de componentes, los medios de comunicación juegan un papel clave, especialmente con respecto al acceso a la agenda política y a la opinión pública.

En un contexto más cercano, así se produjo en los movimientos sociales que nacieron alrededor del 15M. En este caso, desde el punto de vista de la intermediación, no fueron los grandes y tradicionales medios sobre los que pivotó

la construcción de significados, sino ciudadanos y los propios movimientos provistos de las tecnologías de comunicación e información basadas en *Internet*, en especial redes sociales y telefonía móvil.

Aunque existen elementos similares, los marcos de la acción colectiva y los encuadres de los medios mantienen algunas diferencias notables. Para algunos autores, los marcos de acción colectiva son esencialmente ideológicos, mientras que los periodísticos están vinculados por su propia naturaleza a la formulación de significados para ser entendidos por la sociedad. En ambos casos comparten un positivismo o constructivismo de la realidad. Sin embargo, los propios medios de comunicación son actores en la arena política: el propio periodista forma parte de la noticia que describe a modo de hecho objetivo; la enmarca y aporta su particular punto de vista [Git80]. Los medios de comunicación como intermediarios entre la realidad y la sociedad filtran y seleccionan los temas a tratar (al menos en lo relacionado con las propias limitaciones físicas de personas, de tiempo y de espacio).

En la elaboración de noticias informativas a partir de sucesos, los *frames* disponen de dos características fundamentales: reflexividad e indexividad. En la medida en que los relatos que aparecen en las noticias se encuentran dentro de la propia realidad que tratan de describir dándole sentido, los frames invitan a la reflexión. Por su parte, los frames necesitan un contexto, al margen de que se puedan asignar significados fuera de dicho contexto [Sá01]).

También forma parte del *framing* el proceso de elaboración de la noticia. Cobran importancia cómo y por qué se redactan de una determinada forma las informaciones, la relación del periodista con sus fuentes, el mimetismo de los contenidos con respecto a los gabinetes de prensa o relaciones públicas que les proporcionan las notas de prensa, entre otros.

Con la emergencia de las redes sociales se produce una cierta desintermediación en la comunicación política [Ort11], según la cual son los propios actores políticos (partidos, ciudadanos, movimientos sociales e instituciones políticas) los que comienzan a jugar un papel activo en esa comunicación.

Las investigaciones sobre los efectos de los frames se han abordado tratando de identificar tendencias en la definición de cuestiones importantes (*issues*), los distintos tratamiento y cobertura en función del tipo de medio [SV00]. En el ámbito de las noticias, los marcos han recibido la atención para el estudio de campos tan diversos como la acción afirmativa [GM89], el apoyo a la guerra [DKWT05], opiniones sobre la investigación con células madre [NBK03], el cinismo sobre el gobierno [BS02] o las atribuciones de responsabilidad sobre la epidemia de la obesidad [Law04].

Por todo ello, el análisis de frames ofrece claves para entender los procesos de cambio cultural [RL04], los sesgos de los medios [Ent07, TJP+00], la comprensión de lo público [BK06] y la formación de la opinión [CD07]. Se puede afirmar que, en general, los temas nuevos requieren un mayor esfuerzo en la elaboración de marcos, mientras que para tratar los viejos asuntos se suelen utilizar, matizar y alterar marcos de los que se había hecho uso previamente.

Aunque existen ciertas diferencias en cuanto a la utilización de marcos en la acción colectiva como los correspondientes a los medios de comunicación, comparten algunos rasgos:

- Son un instrumento de construcción de significados, por lo que tiene un carácter constructivista de la realidad, situándose en el plano cognoscitivo.
- Están orientados a la acción, y en el caso de los marcos de los partidos, hacia la participación electoral y al voto.
- Tratan de recoger singularidades de la sociedad y otorgarles relevancia política.
- Suponen un vínculo entre los emisores y los receptores, siendo esta vinculación tanto a nivel individual como colectivo. Este vínculo se construye sobre asuntos que generan controversia y identidad política.
- Los marcos son instrumentos de movilización, intrínsecamente ideológicos, difunden y reivindican posturas particulares.
- Ofrecen una determinada visión e interpretación de la realidad, por lo que pueden pasar a construir las agendas públicas o de los medios.

Sin embargo, con respecto a sus efectos, los marcos en las noticias de los medios y los *frames* de correspondientes a los *spots* políticos existen algunas diferencias de consideración. En primer lugar, en ambos tipos no acceden a la información el mismo universo de personas. Mientras que las noticias están accesibles para todos los ciudadanos, los *spots* de campaña en las redes sociales son vistos principalmente por personas con una cierta afinidad partidista o ideológica. Sin embargo, no es menos cierto que el consumo de noticias en los distintos medios también arroja un comportamiento no homogéneo según la identificación partidista [Mei10]. Así, en función de esta aproximación ideológica las personas consultan unos u otros periódicos o televisiones. En el caso de la prensa escrita, la oferta de una heterogeneidad "ideológica" es mayor que en las grandes cadenas de televisión.

2.3.3.2 **Tipos de marcos y el proceso de encuadre**

La importancia de los frames en el proceso comunicativo reside en los efectos que tienen sobre las actitudes y comportamientos [Dru01]. El proceso según el cual los marcos en la comunicación de las elites (políticos, medios, grupos de interés y otros actores políticos) afectan a las actitudes se engloba bajo el término *framing effect*. Chung y Druckman proponen un proceso en tres etapas que tiene lugar en los mecanismos psicológicos [CD07].

Según este modelo, en un primer momento, una determinada consideración o tema debe ser *almacenada* para estar disponible de cara a potencialmente ser recuperada y utilizada en un estadio posterior. En segunda instancia, esta consideración debe ser *accesible* [PT97] para lo que el umbral de esfuerzo para su recuperación debe ser inferior a un cierto nivel (de otro modo, el coste de acceso podría inhibir el poder acceder a él). Una de las formas de proporcionar accesibilidad es mediante una exposición frecuente, repetida, o bien que sea reciente en el tiempo a un marco de comunicación que la enfatice. Como tercera etapa, un individuo evaluará conscientemente la posible aplicabilidad de las consideraciones que están accesibles. Esta aplicabilidad del marco (y por tanto, la probabilidad de que afecte a las opiniones del individuo) es tanto mayor cuanto más intensa o relevante la considera [EC93]. Esta fuerza y relevancia dependen de aspectos semánticos y retóricos del marco: que el *frame* tenga resonancias en la cultura de quien lo recibe, si ofrece información estadística o si es emitido por fuentes creíbles, entre otros aspectos.

A través de trabajos experimentales muestran que la repetibilidad frecuente no es el único componente que afecta a la opinión pública [CD07]. En ocasiones la combinación de distintos *frames* (*framing-mix*), cada uno con distinta relevancia e intensidad, pueden conducir a niveles de eficacia superiores. En estudios recientes se constata que las personas que procesan información online responden más a *frames* anteriores, es decir, sus niveles de respuesta están más influenciados por los marcos previos que por los últimos que reciben en el tiempo. Por el contrario, el efecto de los encuadres recientes es mayor en las personas que procesan la información basándose más en la memoria. Con todo ello, se reasalta la importancia en el modo en que los *frames* se presentan a lo largo del tiempo. En la práctica los *frames* compiten entre sí, y la mayoría de los estudios emplean únicamente uno o dos marcos para representar la cobertura mediática [CD10].

Los efectos sobre el recuerdo tienen consecuencias en la capacidad de respuesta (tener los argumentos accesibles, cercanos – y por ello se necesita que les sean repetidos) de cara a conformar actitudes y comportamientos políticos y, en su caso, electorales o específicos con respecto al voto.

Existe un consenso en la literatura sobre los *frames*, que, en tanto que esquemas interpretativos que permiten a los individuos percibir, organiza y dar sentido a la información que les llega [CD10, Dru01, PK97], generalmente enmarcan las noticias de cuatro formas distintas. A saber: (i) enfatizando el conflicto entre partidos o individuos (el denominado *frame* de conflicto), (ii) centrándose en un individuo como un ejemplo o enfatizando sobre las emociones (*frame* de interés humano), (iii) atribuyendo responsabilidad, dando crédito o criticando a instituciones políticas o individuos (*frame* de responsabilidad) y (iv) poniendo el acento en las consecuencias económicas o económicas que un determinado tema tiene para la audiencia (*frame* de consecuencias económicas).

A continuación se describe con mayor grado de detalle cada uno de estos marcos:

1. **Marco de conflicto o *Conflict frame*.** Este marco de conflicto enfatiza la contienda que se produce entre individuos, grupos o instituciones [Neu92]. Se ha observado que la discusión política entre elites durante las campañas electorales reduce la complejidad sustantiva del debate político hasta un conflicto demasiado simplista [Pat11]. A través del uso de este marco, se plantea el ganar y perder como la principal cuestión, haciendo uso de los lenguajes bélico, de los juegos y de la competición, así como del estilo de competitivo de los individuos y los partidos [Jam93]. Según Neuman los medios trabajan con unos pocos *frames*, y entre ellos el de conflicto es el más utlilizado. Otras investigaciones ha observado que la discusión en las noticias entre las elites a menudo reduce el debate político sustantivo y complejo convirtiéndolos en conflictos muy sencillos, lo que ayuda a simplificar el posicionamiento de la audiencia, hacia posiciones con mucha frecuencia binarias [Neu92] .

2. **Marco de interés humano o *Human interest frame*.** A través de este marco, los eventos, temas o problemáticas se plantean como la historia biográfica de un individuo o desde un enfoque eminentemente emocional. Para ello, trae el lado humano o un ángulo emocional a la presentación de un acontecimiento, temática o problema (del inglés *issue*). Personalizar una problemática contribuye al carácter narrativo de una noticia. El uso de este marco en las noticias es una forma de personalizar, dramatizar y emocionar las noticias. Neumann lo denomina *frame* "de impacto humano", y suele ser frecuentemente utilizado en las noticias de los medios. Es una de las maneras que se utilizan para capturar y retener el interés de la audiencia [Ben95a]. Supone un esfuerzo por personalizar, dramatizar o *"emocionalizar"* las noticias [Neu92].

3. **Marco de responsabilidad o *Responsibility frame*.** Presenta un asunto o problema de modo que le atribuye la responsabilidad de causar o resolverlo al gobierno, a un individuo o a un grupo; y lo hace en el contexto de ciertos principios religiosos o prescripciones morales .

4. **Marco de consecuencias económicas o *Economic consecuences frame*.** Cuando un suceso, problema o temática se plantea en términos de consecuencias económicas para un individuo, grupo, institución, región o país, se está haciendo uso del *frame* de consecuencias económicas. Generalmente se propone en términos del impacto actual o potencial para la audiencia [Neu92]. Según **Iyengar (1991)**, cubrir un asunto en términos de un acontecimiento desde el punto de vista individual más que en términos de un contexto social e histórico, hace que la gente ofrezca explicaciones en clave individual para los problemas sociales.

Aunque no existe un consenso total es sobre si existe un efecto real de estos frames sobre las percepciones de la audiencia. Sí existen evidencias de los efectos de noticias enmarcadas estratégicamente en términos de ganar o perder [CJ96, Iye87, Iye94, PTP97], el acuerdo sí existe en ciertos aspectos. Así, por ejemplo, presentar las cuestiones en términos de eventos, instancias o episódicamente individuales conduce a que la gente ofrezca explicaciones a nivel individual a los cuestiones sociales [Iye94]. Por otra parte, el efecto de las noticias enmarcadas sobre la memoria no ha sido muy estudiado de manera sistemática. Sin embargo, según Graber [Gra90], Gunter [Gun87] y Robison y Levy [DR86], las historias de las noticias están personificadas y cercanas a los temas domésticos se recuerdan mejor. Todo ello muestra que familiarizar las cuestiones resulta eficaz desde el punto de vista comunicacional.

2.3.3.3 El enmarcado negativo

Las investigaciones han mostrado que la interpretación que las personas hacen de la realidad política está influenciada por los marcos en los que se presentan las noticias (o su término anglsajón *news frames*). Algunos trabajos muestran que existe una relación entre la estructura del anuncio político y el proceso psicológico del espectador [Bio91]. Se activan ciertos elementos cognitivos en la memoria para influenciar las actitudes ante las decisiones ya que resulta mucho más eficaz que aquellos elementos cognitivos que han sido frecuentemente o recientemente activados tengan mayo efecto [KS88, Sim57, KST82, Zal92].

Así mismo, ciertos estudios han mostrado que anuncios negativos conducen a crear climas de negatividad política, una gran alienación en el votante y una

devaluación del debate político [Dia92, Gro92, PSH96].

En el estudio de la publicidad negativa, dos son los frames más relevantes [SHPR00, DSG13]. El primero es el de asuntos *ad hoc* (ad hoc issue) y sobre temas del candidato (*candidate theme*):

- **Temas *ad hoc***: Se centran en cuestiones-issues específicas u opciones de políticas públicas que son relevantes en la campaña inmediata, alrededor de los cuales se proponen marcos de interpretación. El anuncio político (en el que se presenta el frame) puede informar sobre eventos que rodean a la cuestión, afirmaciones sobre el candidato oponente o las diferencias en la conducta, las prioridades o las opiniones del otro candidato. Plantea que si el otro candidato es elegido, se producirán consecuencias negativas debido a las posiciones que defiende en uno o más issues-cuestiones.

- **Temas del candidato**: Sobre estos temas, la elección se presenta en términos más amplios, como una competición entre el bien el mal políticos (political good and evil). Se realizan ataques peyorativos, a menudo estereotipados, y los temas se abordan desde la representación, orientación y la personalidad del candidato. Se pretende que se realice una "inferencia" a partir del comportamiento pasado del candidato oponente . En muchas ocasiones se apela a creencias-devociones morales del candidato.

Generalmente se asume que en cierta medida, la magnitud del impacto del frame se debe a la dosis, esto es, la cantidad de publicidad de un frame particular en la memoria de trabajo o memoria a corto plazo [Est84].

La práctica científica respecto a la eficacia de los frames adolece de algunas limitaciones. Así, no existen un conjunto estándar de indicadores que se puedan usar para medir fielmente la prevalencia de marcos comunes en las noticias, permitiendo la comparabilidad [SV00].

2.3.3.4 **La ética como marco**

El uso de la ética y la moral es frecuente en la política estadounidense, donde se ha utilizado en la mayor parte de las últimas campañas electorales [PD09]. Los estudios principalmente de Lakoff sobre el perfil del ideal tipo votante demócrata y republicano ofrecen una metáfora basada en la figura del padre de familia. Así, un votante republicano respondería a las aspiraciones de el padre autoritario más exigente, racional, y que demanda rigor y disciplina. En el otro lado, un votante demócrata pretendería un padre más protector, comprensivo y facilitador [Lak02].

Desde el punto de vista moral, dada la relevancia del papel del candidato en los procesos electorales, se han venido presentando esta figura desde distintos estereotipos. Es a través de estos estereotipos sobre los que el marco de moralidad muestra una apreciable eficacia. A continuación se enumeran y describen las distintas formulaciones que constituyen el marco de moralidad [Fos04, Bar96].

1. **Inmoralidad del oponente**. El marco moral presenta un candidato inmoral desde distintos puntos de vista, a saber:

    - *Fraude demostrado del oponente*. Mediante el empleo de esta formulación se explicitan evidencias de que el adversario ha cometido.
    - *El adversario provoca daño*. Se presentan las actuaciones del contrincante político como causantes de un perjuicio a terceros.
    - *Identificar al contrincante como enemigo del pueblo*. Se elabora una construcción comunicacional mediante la cual se sitúa a una persona o colectivo antagónico como amenaza o peligro para el pueblo como sociedad, de cara a su supervivencia o a la consecución de alguno de sus objetivos clave.
    - *El oponente es un líder fallido*. Según esta metáfora, el adversario no estaría legitimado en tanto en cuanto no dispone de capacidad de tracción o poder dentro de su propia organización.

2. **Candidato populista**. Atendiendo a este planteamiento, el candidato está alineado con la defensa de los intereses generales de una sociedad. En este punto conviene señalar que el término *populismo* ha venido adquiriendo un matiz peyorativo, aplicándose como sinónimo de demagogia [GV11]. Al margen de esta consideración, un candidato sería presentado como populista cuando dispone de alguna de estas cualidades:

    - *Es confiable y honesto*. Se construye la idea de que el candidato es una persona en la que se puede confiar.
    - *Es independiente y justo*. Según esta idea, el candidato mantiene ecuanimidad en sus posiciones y planteamientos.
    - *Comparte los valores del país o territorio*. El marco establece una correspondencia clara entre los valores del territorio y los del candidato.

3. **El candidato como héroe**. La persona que representa al partido en calidad de máximo responsable o de candidato es identificado como héroe. Esta heroicidad se puede expresar en los siguientes términos:

- *Batalla contra los enemigos*. Mediante esta propuesta, el candidato defiende los intereses de un colectivo frente a terceros.
- *Actúa como protector del pueblo*. A través de este símil se plantea que el candidato fija como prioridad el bienestar y progreso de la sociedad a la que representa.
- *Es firme y trabajador*. Con este marco se pone de relieve el rigor ético y el esfuerzo del candidato.

4. **El candidato como estratega**. Este enfoque opera cuando el candidato se presenta de alguna de las siguientes maneras:
   - *Progresista*. Se presenta al candidato como una persona orientada hacia el futuro.
   - *Pragmático*. El candidato aparece como orientado principalmente a los resultados.
5. **El candidato como víctima**. Al hacer uso de este marco, el candidato juega el papel de perdedor dentro de un proceso de pugna entre instituciones, partidos políticos o incluso dentro de un partido político.
   - Acusado erróneamente. El candidato es presentado como la víctima de una falsa acusación de la que ha sido víctima con o sin intencionalidad.
   - Reputación dañada. Cuando se presenta el prestigio social del candidato propio.

2.3.3.5 **El proceso de alineamiento de marcos**

Para que un marco determinado tenga eficacia comunicativa es necesario que exista una cierta asimilación entre el marco propuesto como estímulo y la respuesta que ello provoca. Este proceso, denominado *alineamiento de frames*, es la conexión entre las orientaciones de un colectivo origen y las individuales de una determinada persona, tales como los intereses individuales, valores y creencias, de modo que sean congruentes y complementarios con las actividades, las metas y la ideología de un grupo origen, sea este una ONG, un movimiento social, un sindicato o u partido político [SRJWB86]. De este modo, la alineación de marcos es una condición necesaria para la identificación a un determinado colectivo, la adhesión a sus principios, o la participación en un movimiento social, independientemente de su naturaleza o intensidad.

Existen distintos tipos de procesos de alineamientos de marcos

1. **Conexión de marcos o *frame bridging***. Se produce este tipo de conexión cuando se vinculan marcos de dos más marcos que son ideológicamente congruentes, pero estructuralmente desconectados con respecto a una determinada cuestión (*issue*). Puede suceder a nivel organizativo o a nivel individual. Son la vinculación de un actor político con lo que McCarthy [McC86] denominó como *pools* de sentimientos no movilizados o *clusters* de preferencias de la opinión pública, que agregan a personas que comparten sentimiento de agravio pero carecen de la base organizativa para expresar su descontento y actuar para alcanzar sus objetivos. En estos casos, la acción colectiva no es precedida por la toma de consciencia o la transformación de marcos.

   Se lleva a cabo, en primer lugar, por medio de la participación comprometida (*engagement*) en la organización, y además por la difusión de información a través de redes interpersonales o intergrupo, los medios, el teléfono, el correo y en los últimos años mediante las TICs, filtrando y seleccionando posibles listas de destinatarios. El frame bridging abunda en los movimientos sociales contemporáneos.

2. **Amplificación de marcos o *frame amplification*** Consiste en clarificar o reforzar un marco de interpretación que tata sobre un tema, problema o conjunto de sucesos. Dado que el significado de los sucesos está muchas veces envuelto en indiferencia, decepción, o por la ambigüedad o la incertidumbre [Gof74] es necesaria la amplificación. Existen dos modelos de amplificación de marcos:

   (a) Amplificación de valores:

   - Los valores se construyen como modos de conducta o estados de existencia que una persona cree que merece la pena proteger y promocionar [Rok68, RK72]. En este sentido, los valores se ordenan jerárquicamente, dando preeminencia a unos sobre otros [Rok68, Wil70].
   - La amplificación de los valores consiste en la identificación, idealización y elevación de uno o más valores que se presumen como fundamentales pero que no han sido considerados como potencialmente constitutivos por cualquier otra razón - atrofiados, caídos en desuso, suprimidos por falta de oportunidad para ser expresados [Til78], porque se han dado por sentados o garantizados [Zij79], no suficientemente desaiafos o amenazados [TK57] , o porque su relevancia en un determinado suceso o problemática-issue puede resultar ambigua [Gof74].

- Entre los valores más amplificados por los actores políticos durante las últimas décadas los están ligados a la democracia igualdad y libertad. Así, tradicionalmente han sido la familia, la etnicidad, la propiedad, la vecindad y la integridad. En los últimos años se han incorporado la justifica, la cooperación, la perseverancia, y el carácter sagrado de la vida humana [SRJWB86].

(b) Amplificación de creencias:

- Las creencias dan soporte cognitivo para llevar a cabo o impedir tomar acciones para logar un determinado valor.
- Las creencias se refieren a asociaciones que se presume que existen entre dos cosas o entre una cosa y una característica de ella [Bem70]. Según este modelo, existen cinco tipos de creencias: (i) acerca de la seriedad de un problema, cuestión o agravio en cuestión; (ii) sobre el origen de la causa o culpa; (iii) creencias estereotipadas sobre antagonistas o destinatarios de la influencia; (iv) sobre la probabilidad de cambio o la eficacia de la acción colectiva; y (v) Sobre la necesidad y el decoro de mantenerse firme y/o resistir.

2.4 Métodos de procesado de datos y textos

El descubrimiento de conocimiento o *Knowledge Discovery* pretende identificar patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles en los datos [FPSS$^+$96]. Para la obtención de este conocimiento oculto en los datos es necesario aplicar un proceso por fases y repetitivo. En ocasiones estos pasos requieren interacción humana, por lo que no son totalmente automáticos.

El grado de eficacia de estas técnicas se puede medir en términos de distintas características. Así, algunas de ellas son la inteligibilidad para las personas, la validez en el contexto de las medidas estadísticas dadas, la novedad de ese conocimiento y la utilidad.

2.4.1 El descubrimiento de conocimiento

El descubrimiento de conocimiento (KD, del inglés *Knowledge Discovery*), o descubrimiento de conocimiento en bases de datos (KDD, de los términos ingleses *Knowledge Discovery Databases*) se define como «el proceso no trivial de identificación de patrones válidos, innovadores, potencialmente útiles y en último lugar, comprensibles, en los datos» [FPSS$^+$96].

En general, el proceso de descubrimiento de conocimiento pretende identificar determinados patrones en los datos, de modo que sean característicos de ciertos comportamientos y las posibles relaciones entre ellos. Por consiguiente, estos procesos requieren gran esfuerzo y gran conocimiento de un experto humano, en ocasiones incluyendo una interacción directa por su parte.

Durante el proceso de extracción de conocimiento a partir de conjuntos de datos, resulta necesario establecer unos criterios que permitan distinguir distintos grados en la calidad del conocimiento encontrado. Así, es muy frecuente el uso de la utilidad aportada, la capacidad innovadora del patrón, la validez o representatividad estadística en su contexto, así como la facilidad de ser interpretado de la manera más sencilla para el ser humano.

2.4.2 La minería de datos

El concepto de minería de datos (de la voz anglosajona *data mining*) comporta distintas acepciones. La primera de ellas, es la que la asimila con el término descubrimiento de conocimiento, mientras que en una segunda interpretación, equivale a una de las etapas del proceso de descubrimiento de conocimiento en lo que se refiere a la creación de modelos que representan conocimiento [FPSS+96].

Se trata de un ámbito –el de la minería de datos– en el que los cambios son continuos y profundos, proporcionando notables cotas de avance en su formulación y contribución científica [HK06], desde distintas perspectivas y con un planteamiento transdisciplinar.

Cabe señalar que la representación de conocimiento y su posterior extracción a partir de unas fuentes dadas requiere de capacidades e infraestructura. Es por ello que el avance de la minería de datos ha venido desarrollándose en las últimas décadas por el progreso en tres áreas con las que está directamente vinculado: la estadística, las bases de datos, las técnicas de aprendizaje automático. Y ello ha sido posible gracias al avance en los desarrollos *software* y *hardware* sobre los que se han construido herramientas y tecnologías.

- **Estadística**. El sustrato de la minería de datos es de naturaleza matemática, por lo que la estadística y su relación con el álgebra son de especial importancia para poder disponer de una representación homogénea y consistente de los datos. Gran parte de los métodos que forman parte del cuerpo de la minería son estadísticos, como bien quedan recogidos en [FHT01, BH00, Mai02].

- **Bases de datos**. Los sistemas de almacenamiento de información requieren, no solo disponer de mecanismos para la custodia, acceso rápido y capacidad para manipular los datos, sino prestaciones relacionadas con una rápida capacidad operativa mediante técnicas algorítmicas más o menos sofisticadas de minería de datos. Su papel es vital para poder procesar esa información en un tiempo razonable y con el menor coste posible [CHY96].

- **Aprendizaje automático**. El aprendizaje automático o, en inglés, *Machine Learning* es un conjunto de técnicas orientadas hacia el proceso de aprendizaje por parte de los ordenadores en base a unos datos de partida, generalmente simbólicos. Dada la heterogeneidad de estos datos y la dificultad intrínseca de dicha heterogeneidad, el aprendizaje automático necesita analizar la complejidad algorítmica de las distintas implementaciones físicas que se construyen para poder producir dicho aprendizaje. Gran parte de las técnicas que se utilizan frecuentemente quedan recogidas en [Mit97].

2.4.3 La minería de textos

La Web es un agente *habilitador*, tecnológicamente hablando, que fomenta la creación de una gran cantidad de contenido textual por diferentes usuarios, gracias a la facilidad y sencillez con la que se guarda y procesa este contenido textual. La creciente cantidad de datos de textos disponible en diversas aplicaciones ha creado la necesidad de realizar avances en el diseño de algoritmos que sean capaces de aprender patrones de los datos de una manera dinámica y escalable.

Mientras los datos estructurados son manejados por sistemas de bases de datos, generalmente, los datos de textos son gestionados a través de un motor de búsqueda, debido a la falta de estructuras [CMS10]. Un motor de búsqueda permite al usuario encontrar información útil en una colección con una consulta de palabras clave. Cómo mejorar la efectividad y eficiencia del propio motor ha sido uno de los temas centrales a investigar en el campo de la recuperación de información [BYRN$^+$99a, Jon97].

Sin embargo, tradicionalmente la investigación sobre recuperación de la información se ha centrado más en facilitar el acceso a la información [Jon97] en lugar de analizar la información para descubrir patrones, los cuales son el objetivo principal de la minería de textos.

La minería de textos, también denominada análisis de textos o descubrimien-

to de conocimiento a partir de texto (KDT, de los términos ingleses *Knowledge Discovery from Text*), mencionado por primera vez por [FD95], consiste en el análisis automático de textos empleando técnicas de:

- Recuperación de información (IR, del inglés *Information Retrieval*).
- Extracción de información (IE, de las voces inglesas *Information Extraction*).
- Procesado de lenguaje natural (NLP, de los términos ingleses *Natural Language Processing*).
- Minería de datos (ML o *Machine Learning* en inglés)

Tradicionalmente, el esquema de extracción de conocimiento a partir de texto es el siguiente, que se muestra en la Figura 2.1

1. *Preprocesado de textos*. En esta etapa se trata de hacer que los documentos sean representados en texto de una manera más nítida y consistente. Generalmente esto implica métodos de parada y eliminación de palabras (del inglés *stop word removal*) y *lematización*, esto es, la extracción de las raíz de las palabras.
2. *Representación de texto*. Esta representación suele realizarse en forma de *Bag of Words* o de un modelo de espacio vectorial (VSM), del inglés *Vector Space Model*, formando una estructura de representación de la inforamción. De esta forma es posible transformar el texto en estructuras numéricas que lo representan y con las cuales es posible realizar operaciones.
3. *Descubrimiento de información*. Convertido el texto en vectores numéricos se les puede aplicar métodos de aprendizaje máquina o *machine learning*, y minería de datos tales como clasificación y extracción de conglomerados o *clustering*.

Estas técnicas son explicadas posteriormente en la sección 2.5. A su vez, la minería de texto también tiene conexión con los algoritmos y métodos anteriormente expresados en las secciones 2.4.1 y 2.4.2:

- El descubrimiento de conocimiento en bases de datos.
- La minería de datos.
- El aprendizaje automático.

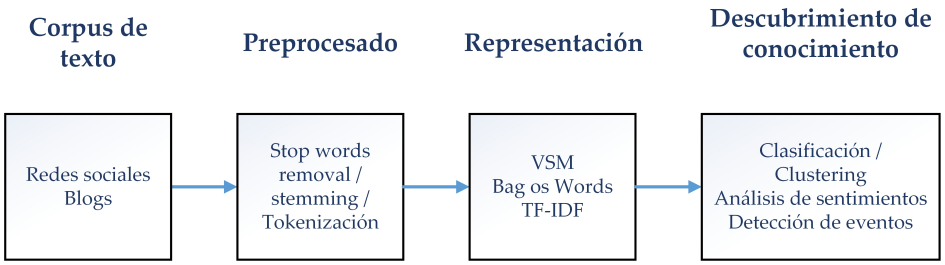


Figura 2.1: Framework clásico de la minería de textos.

- La estadística.

Un aspecto importante de la minería de textos es su continua actualización e innovación, ya que existen diversos grupos de investigadores que trabajan en los nuevos aspectos de la minería de textos, en el contexto de plataformas emergentes como los flujos de datos y las redes sociales.

2.5 Áreas de investigación relacionadas

En la minería de textos, la adecuada selección de características, la influencia que tiene conocimiento experto del dominio en cuestión y de sus técnicas tiene gran relevancia a la hora de conducir a buenos resultados. En particular, la relación bidireccional entre minería de datos y minería de textos implica que sean necesarias determinadas adaptaciones o innovaciones que permitan aplicar técnicas de un campo en el otro y viceversa. Para ello, nos nutriremos de la experiencia y los resultados de investigación que existen en tres ámbitos, como son recuperación de información, el procesamiento del lenguaje natural y la extracción de información.

2.5.0.1 La recuperación de información

La recuperación de información (IR, del inglés *Information Retrieval*) es en un sentido más amplio, se ocupa de toda la gama de procesamiento de información, desde la recuperación de datos hasta la recuperación de conocimiento [Jon97].

La acepción más generalizada es la que tiene que ver con la obtención de documentos que contienen las respuestas a las preguntas realizadas. Esto no se debe confundir con la simple consecución de respuestas por sí mismas [Hea99]. Con este propósito, se utilizan medidas y métodos estadísticos que permiten el procesamiento automático de datos de texto, y también la comparación con la pregunta realizada.

La recuperación de información es una área de investigación con varias décadas de bagaje, donde los primeros intentos para la indexación automática de documentos se hicieron en 1975 [SWY75]. En cambio, el desarrollo de *Internet*, y en particular de la *WWW* o *World Wide Web*) y la necesidad de motores de búsqueda más sofisticados propició su despegue de manera notable.

Aunque la definición de recuperación de la información se basa en la idea de preguntas y respuestas, los sistemas que recuperan documentos basados en palabras clave, es decir, los sistemas que llevan a cabo la *recuperación de documentos* como la mayoría de los motores de búsqueda, son frecuentemente denominados sistemas de recuperación de información.

2.5.0.2 **El procesamiento del lenguaje natural**

El área que estudia los problemas relacionados con el procesamiento y manipulación de los lenguajes naturales se denomina Procesamiento del Lenguaje Natural (o NLP, del término anglosajón *Natural Language Processing*). De manera resumida, el NLP trata de adquirir el conocimiento sobre cómo las personas entienden y utilizan el lenguaje para, con ello, poder construir técnicas y herramientas que puedan ser utilizadas para que otras aplicaciones informáticas sean capaces de entender y gestionar dicho conocimiento.

Resumidamente, el objetivo general del procesado de lenguaje natural es lograr una mejor comprensión del propio lenguaje natural mediante el uso de ordenadores [Kod99]. Otros autores incluyen también el empleo de técnicas sencillas y duraderas para el procesado rápido de texto, tal como presentan [BAT91]. Sus bases habitan en una amplia serie de disciplinas: ciencias de la información, matemáticas, lingüística, inteligencia artificial, robótica, psicología, etc.

El rango de técnicas utilizadas se extiende desde la simple manipulación de cadenas de texto hasta el tratamiento automatizado de consultas en lenguaje natural. Además, las técnicas de análisis lingüístico se utilizan, entre otras cosas, para el procesamiento de texto en lenguaje natural, interfaces de usuario y reconocimiento de voz, entre otras.

2.5.0.3 **La extracción de información**

La extracción de información o IE, de los términos anglosajones *Information Extraction* es una parte del procesamiento del lenguaje natural que trata de obtener información semántica que se puede describir previamente. Para ello es necesario describir la información que se desea extraer, asignándole una deter-

minada caracterización mediante atributos, lo que facilita su descripción y por lo tanto su acceso. De ese modo es posible realizar una selección o filtrado de la información que es relevante en cada caso, evitando de este modo un exceso de contenido informacional presente en el lenguaje natural. Esto los diferencia de los métodos de extracción en documentos de texto, cuya información se procesa y registra en sistemas de bases de datos con patrones similares [Wil97] para su utilización posterior.

La tarea de extracción de información se descompone en una serie de etapas de procesamiento, que incluyen:

- La *tokenización*[1].
- La segmentación de la oración.
- La asignación de *parte de la oración* (POS, de las voces inglesas *Part Of Speech*)[2].
- La identificación de entidades específicas; es decir, nombres de personas, nombres de lugares y nombres de organizaciones.

Algunas frases y oraciones de mayor nivel tienen que ser analizadas, semánticamente interpretadas e integradas. Aunque los sistemas de información más precisos de extracción a menudo implican módulos manuales de procesado de lenguaje, se han logrado progresos en la aplicación de técnicas de minería de datos a este proceso.

2.6 Métodos para el filtrado de contenidos y textos

2.6.1 El filtrado de contenidos

La proliferación de contenidos, fuentes de información, así como la heterogeneidad de dichos contenidos constituye un desafío a la hora de que los usuarios accedan a la información adecuada que demandan, necesitan o esperan encontrar. Entre otras razones, el proceso de filtrado puede estar condicionado por razones de seguridad o privacidad, idoneidad o selección según distintos criterios o simplemente para encontrar los contenidos exactos a partir de una

[1] *Tokenizar* es la acción de dividir un flujo de texto en palabras, frases, símbolos u otros elementos denominados *tokens*.

[2] El *Part Of Speech* es la categorización de las palabras según su comportamiento sintáctico o morfológico.

cantidad creciente de información. Para abordar esta cuestión, se han venido desarrollando sistemas de filtrado que han abordado distintas estrategias debido a la dificultad en el tratamiento de la información, y ante la expectativa de un desarrollo mayor de la futura web semántica[1].

El crecimiento en el uso de *Internet* en los últimos años ha provocado que el proceso de filtrado llegue a operar en distintos niveles y servicios de la arquitectura TCP/IP (*Transmission Control Protocol/Internet Protocol*) [2]. Gracias a este desarrollo surge la necesidad de regular y controlar el tráfico de red, y en consecuencia la de impedir el acceso, transmisión y entrega de información no deseada bajo determinadas circunstancias. El filtrado de *spam* y el filtrado de datos red son casos paradigmáticos de ámbitos en los que este tipo de técnicas.

En relación con la información *online*, se plantea la dificultad de que ésta se presenta en distintos formatos (textos, imagen, audio y vídeo) por lo que se requiere un tratamiento específico para cada uno de ellos para extraer el valor semántico de cada uno de ellos. Pese a que se han producido avances en este área, las técnicas actuales no permiten una representación precisa de dichos contenidos multimedia. En el caso de los motores de búsqueda, la cantidad de información requerida y transferida en cada consulta es extremadamente elevada en relación al resultado, lo que reduce la calidad de este tipo de servicios.

Esta complejidad se acentúa cuando los usuarios tratan de acceder a información o recursos que cumplen unas determinadas condiciones de partida teniendo que considerar aspectos relacionados con la seguridad, la privacidad o la idoneidad de los mismos descrita según diferentes propiedades. En tal circunstancia, el proceso de filtrado requiere una descripción mucho mayor que en condiciones estándar a la hora de proporcionar los resultados esperados y según determinados términos de calidad.

Dada la complejidad a la hora de abordar estos procesos, los sistemas de filtrado sobre web abordan dos cuestiones fundamentales: la protección y el rendimiento. En primer lugar, se trata de garantizar unos esquemas mínimos de protección que eviten acceder a información peligrosa o perjudicial, tratando de minimizar los perniciosos efectos que se podrían ocasionar en caso contrario.

[1] A través de la incorporación de información descriptiva sobre el significado de la información, la web semántica supone un nuevo marco de presentación y acceso a la información y servicios en *Internet*. Para ello se hace necesario añadir descripciones explícitas del significado de los datos, de su estructura interna y de la estructura global de los servicios y la información disponibles en la WWW [BLHL+01].

[2] La arquitectura TCP/IP permite a cualquier clase de computador comunicarse entre sí [SW94b] a través de una red. Esta arquitectura se construye en base a 4 capas que establecen relaciones entre sí: (i) capa de enlace, (ii) capa de red, (iii) capa de transporte y (iv) capa de aplicación.

En segundo término, se pretende ofrecer una respuesta en un plazo de tiempo razonable entre el envío de la consulta, de modo que se minimice el retardo. En todo caso, se procura un equilibrio entre el tiempo de respuesta y la calidad de la respuesta ofrecida.

2.6.1.1 **Enfoques de filtrado web y valoración**

Dado el auge de la web durante los últimos años, la disponibilidad de información y servicios se enfrenta, tanto a un incremento en la demanda de acceso a ellos, como a riesgos derivados del acceso ilícito o ilegítmo a dichos contenidos, además de la vulneración de derechos de protección de la intimidad y otros aspectos de seguridad, tanto de la información como de las personas.

Los mecanismos de protección pueden operar a distintos niveles de la arquitectura TCP/IP, son agentes especializados en esta materia los que operan ofreciendo servicios que se basan en el filtrado. Estos agentes pueden ser proveedores de servicios de *Internet*, (*ISP* o *Internet Service Provider*), empresas TIC especializadas y determinadas agencias gubernamentales o privadas, que se ocupan de la valoración y filtrado de la información antes de hacerla llegar a los usuarios finales de la misma. Las políticas, herramientas y técnicas empleadas están en función del tipo concreto de servicio que se contrate y la misión o encargo que dichas agencias o empresas tengan asignado.

En términos operativos, la estrategia más generalizada consiste en una valoración manual o automatizada de ciertos contenidos web, para lo que se emplean distintas técnicas, como las basadas en redes neuronales, para posteriormente asignarles un valor dentro de un conjunto de categorías predefinido. De este modo, es posible tener vinculado cada uno de los tipos de contenido con distintas categorías que indican los tipos de acceso están permitidos y cuáles no, y así se posibilita que los clientes finalmente puedan llegar a aquellos contenidos a los que deberían poder acceder. Un mecanismo similar es el utilizado por el motor del buscador Google para configurar las políticas de acceso a ciertos contenidos considerados inadecuados mediante la utilidad *Google SafeSearch* [1].

La valoración de los contenidos de la web es una etapa necesaria para el posterior filtrado de los contenidos. Generalmente se realiza de manera estática, previamente a cada una de las peticiones de acceso al contenido. De este modo se reduce el tiempo de carga entre la consulta y la respuesta, aunque requiere un esfuerzo de valoración Esta valoración supone una clasificación previa de dichos contenidos, y se puede realizar desde dos enfoques. El primero de ellos consiste en asignarle un valor o categoría (que se puede definir previamente) a cargo de

[1] `http://support.google.com/websearch/answer/510`

terceras personas u organizaciones ajenas al proveedor del contenido o a quien lo ofrece a través de *Internet*, mientras que el segundo, la autovaloración, es realizado por el propio sitio web de manera más o menos automatizada.

La evolución de los métodos de filtrado ha conducido a que los avances en este campo mejoren y flexibilicen las técnicas preexistentes a través de la aplicación de dos principios fundamentales [BFP06]:

1. El método de filtrado ha de ser válido para distintas técnicas de valoración y filtrado (ya sean directas o indirectas), por lo que pueden ser utilizadas individualmente o de manera combinada, según las necesidades de los usuarios.

2. Para una eficacia en las prestaciones, el método de filtrado ha de incluir una descripción precisa de los usuarios en términos de atributos que permitan una mejor caracterización de los mismos.

Tanto los recursos a los que se permite el acceso, como los individuos que tratan de acceder a ellos se describen de manera explícita o implícita, de modo que sirve para la estrategia de filtrado. Así, por ejemplo, las especificaciones explícitas corresponden a los enfoques basados en listas (es decir, listas blancas o listas negras que recogen el conjunto de personas que sí pueden acceder al recurso y aquellas que lo tienen prohibido o limitado respectivamente). En cambio, las especificaciones implícitas combinan todas las estrategias que se basan en las calificaciones, como por ejemplo: PICS (acrónimo del inglés *Platform for Internet Content Selection* [RM96]) y / o categorías de contenido (por ejemplo, *RuleSpace*, una herramienta adoptada por *Yahoo*[1] y otros ISPs para sus servicios de control parental[2]). Esta integración de especificaciones del sistema extrínseco difiere totalmente de enfoques como el basado en el modelo ABC [LH01], con el objetivo de proporcionar la interoperabilidad semántica entre ontologías.

El modelo formal para el filtrado web, *MSFM*, de la expresión anglosajona *Multi-Strategy Filtering Model* establece un marco de trabajo para el de filtrado, que incorpora materias como la flexibilidad y la protección. Incluye capacidad de personalización en cierto grado, para lo cual hace uso únicamente de una parte limitada de los atributos con los que se caracteriza a los usuarios y a los recursos. En este marco, se describen los atributos de los ambos tipos de entidades -usuarios y recursos- de modo explícito mediante identificadores o bien

[1] `http://www.yahoo.com/`
[2] `http://www.symantec.com/popup.jsp?popupid=parental_controls`

de modo implícito incorporando las restricciones en sus propiedades en sentido negativo.

2.6.2 El filtrado de textos

El filtrado de textos tiene un objetivo en sí mismo, que no es otro que facilitar a un usuario el acceso adecuado en determinadas condiciones a un recurso. Sin embargo, contribuye al avance de otras áreas de investigación como son la recuperación de información, el modelado de usuarios, los sistemas de recomendación o el aprendizaje automático entre otras áreas.

Los sistemas que se construyen con tal fin han de estar provistos de capacidades para identificar similitudes entre documentos de texto y así ser capaces de proporcionar mecanismos de comparación entre documentos. Gran parte de las investigaciones sobre recuperación de textos hacen uso de la frecuencia con la que determinadas partes de ese texto aparecen en el documento sobre el que se realiza la búsqueda. En paralelo, los sistemas de recomendación han venido evolucionando añadiendo un conjunto de características adicionales que están basadas en anotaciones compartidas provenientes de otros usuarios.

La convergencia de los trabajos sobre recuperación de textos y los estudios acerca de los sistemas de recomendación ha supuesto un notable avance [Oar97] en la medida en que mejora el rendimiento particular de cada uno de ambos enfoques en tanto en cuanto realiza búsquedas sobre documentos filtrados al mismo tiempo que permite definir con mayor nitidez las características o *perfil* del usuario.

El trabajo existente en la retroalimentación implícita muestra esperanza, y mediante el aprovechamiento de esa promesa puede ser posible mejorar significativamente el rendimiento del componente de cooperación de un sistema integrado de filtrado de textos. Con una base fértil para el futuro desarrollo, el autor asegura que los futuros sistemas de filtrado de textos explorarán aspectos interesantes y encontrarán aplicaciones en áreas que son críticas, para el uso eficaz de la infraestructura de información global emergente.

A partir de un conjunto desestructurado de textos, la clasificación del textos elabora una estructura que permite un almacenamiento, búsqueda y navegación más fáciles, directos y rápido [Seb06]. A este fin se establece una serie de actividades que tratan de abordar estas fases, principalmente relacionadas con dos cuestiones. En primer lugar, el tipo de aprendizaje automático que se utiliza para la clasificación, y que ha sido explicado con detenimiento en la sección 4.1) y, en segundo lugar, la granularidad de los elementos del texto [GSGR09].

- Los 2 tipos de aprendizaje son definidos por el conjunto de entrenamiento de control:

    1. **Aprendizaje supervisado**. El conjunto de clases se conoce previamente a construir el conjunto de entrenamiento, y además se dispone de instancias para cada una de las clases.
    2. **Aprendizaje no supervisado o (*clustering*)**. El conjunto de clases no se conoce antes del entrenamiento, por lo que la meta consiste en agrupar entidades textuales en relación a un contenido similar.

- Con respecto a la granularidad, se pueden definir 3 niveles: términos o *tokens*, frases o documentos:

    1. **Términos**. Incluyen desde la raíz de la palabra hasta palabras sueltas o las expresiones cortas.
    2. **Frases**. Abarcan desde las oraciones hasta las frases complejas.
    3. **Documentos**. Son escritos de tamaño reducido, medio o con un gran volumen de texto.

Estas actividades pueden tener peso por sí mismas o formar parte de procesamientos mayores, tales como el etiquetado de partes del discurso o *POS* (del término anglosajón *Part of Speech*), el reconocimiento de entidades nombradas o la desambiguación de términos, que con notable frecuencia son empleadas como etapas anteriores a la categorización de textos con el fin de aumentar la calidad de los atributos mediante la asignación de un valor semántico adecuado a cada término.

Actualmente, existen diferentes modelos para poder categorizar y filtrar textos en *Internet*. Como por ejemplo en [HSGdBR03], en donde la representación más utilizada es el modelo de espacio vectorial (*VSM*, de los términos ingleses *Vector Space Model*), a la vez que aplica una lista de palabras vacía, cuando el lenguaje puede soportarlo (normalmente lenguajes occidentales). Por otra parte, también aplican diferentes técnicas como la ganancia de información (IG, de las voces inglesas *Information Gain*) que se utiliza para filtrar aquellos términos que no parecen ser importantes para la detección, así como diferentes algoritmos de aprendizaje automático como: el clasificador bayesiano ingenuo (del inglés *Naïve Bayes*), los árboles de decisión (de los términos ingleses *Decision Trees*) o las máquinas de soporte vectorial (SVM, acrónimo del inglés *Support Vector Machines*). Todo este proceso se desarrolla más profundamente en los siguiente capítulos.

2.7 **Sumario**

A lo largo de este capítulo hemos abordado una revisión de la literatura sobre los temas centrales sobre los que pivota la investigación.

En primer lugar, hemos constatado que el desarrollo de *Internet* como instrumento tecnológico supera los propios límites de la tecnología y como consecuencia ha afectado a los procesos sociales y políticos. En general, todos los actores del sistema político dialogan bidireccionalmente con la tecnología. Por su parte, los partidos políticos destacan por el desarrollo de instrumentos basados en *Internet* para poder comunicarse con la sociedad y así poder contribuir a un desarrollo democrático de las sociedades a las que representan.

Hemos visto, en segundo término, cómo la obtención de representación institucional por parte de los partidos supone una competición con el resto de opciones. Por este motivo se han presentado las principales Escuelas que estudian y explican el voto de los ciudadanos, desde distintas perspectivas y con implicaciones completamente diferentes. Así mismo, se han presentado los procesos electorales y uno de los principales elementos de este tipo de procesos, como son las campañas electorales.

El capítulo ha continuado conociendo la importancia del factor comunicacional en la política, y en especial durante las elecciones, rastreando los principales cambios de la comunicación política durante las últimas décadas hasta la actualidad. En este punto, han emergido elementos novedosos con respecto a los clásicos, como son *Internet* o las más recientes *sitios web de interacción social* o simplemente redes sociales digitales. Desde el estudio de la comunicación nos hemos acercado al análisis de marcos como una forma de reconocer e identificar elementos estructurales y estratégicos de la comunicación que cuentan con un notable desarrollo teórico junto a un enorme bagaje práctico en la arena política.

En la última parte del capítulo se ha repasado el estado de la cuestión sobre los distintos enfoques y métodos del procesado de datos y textos. Sobre la base del descubrimiento de conocimiento hemos explorado la minería de datos y la minería de textos. Hemos repasado la vinculación de estas dos áreas con otras de interés para la investigación como son la recuperación de información, el procesamiento del lenguaje natural y la extracción de información. De igual modo, nos hemos acercado al filtrado de contenidos y el filtrado de textos como casos de métodos para el filtrado de datos. También hemos revisado diferentes métodos de filtrado de textos que serán de utilidad a lo largo de la presente investigación doctoral.

«Hay que admitir que el poder produce saber;
que poder y saber se implican directamente uno al otro;
que no existe relación de poder
sin constitución correlativa de un campo de saber,
ni de saber que no suponga y no constituya al mismo tiempo
relaciones de poder.»

Michel Foucault. *Vigilar y Castigar* (1926-1984)

CAPÍTULO

3

Modelado y procesado inicial de los datos

DURANTE este capítulo se explican métodos y técnicas que se han utilizado para modelizar toda la información que se ha recopilado durante el proceso de captura de datos desde la plataforma *Facebook*, así como las etapas posteriores que conducen a la obtención de resultados. Esta etapa permite realizar el proceso de entrenamiento y aprendizaje de los modelos que representan a las conversaciones y los mensajes políticos, de modo que puedan ser probados empíricamente en los capítulos posteriores.

En las siguientes páginas se describe cómo los datos se han recogido de manera automática, para un posterior filtrado que ha permitido comprobar la veracidad y la calidad de los mismos. Tras ello, durante esta fase también se ha realizado el etiquetado manual de los conjuntos de datos. Una vez seleccionados éstos, se han representado según el modelo de espacio vectorial (VSM, del inglés *Vector Space Model*), de modo que han podido ser tratados de manera adecuada.

El capítulo se estructura según las tareas llevadas a cabo durante esta fase. En primer lugar, en la sección 3.1 se presenta la arquitectura general del sistema propuesto para, en la sección 3.2 describir el sitio de interacción social *Facebook* sobre el que se han realizado la extracción de los datos para la experimentación. Este proceso de captura de información se describe con detalle en la sección 3.3.

Posteriormente, la tarea de etiquetado de los distintos conjuntos de datos

se recoge en la sección 3.4, que incluye en su parte final el análisis de los resultados de dicho etiquetado. Continuando con el proceso, en la sección 3.5 se incluye la adaptación de los datos. En esta sección se describe el proceso de extracción de las características relevantes de los conjuntos de datos de mensajes y conversaciones. Por último, la sección 3.6 resume los conceptos expresados a lo largo de este capítulo.

3.1 Arquitectura de la solución propuesta

Para el presente estudio, se ha de desarrollar una serie de fases que permitirán ir alcanzando diversas metas parciales. A partir del diseño de la investigación explicitado en el capítulo 1 se ha elaborado una arquitectura de la solución, recogida en la Figura 3.1.

1. Obtención de información de perfiles de partidos políticos. Análisis descriptivo del mapa digital de la presencia de partidos en redes sociales, y sus vínculos sociales. Tras la selección de los perfiles políticos de los partidos en cuestión, es necesario obtener los mensajes políticos que se han emitido dentro de las sucesivas conversaciones o diálogos políticos en dichas plataformas sociales, así como la información relevante de cada uno de ellos.

2. Construcción de un modelo de aprendizaje para la clasificación de los mensajes políticos sobre el que elaborar el posterior sistema de clasificación.

3. Construcción de un sistema de clasificación de conversaciones políticas que permite, de manera automática, reconocer determinadas características relevantes de las conversaciones iniciadas por los partidos políticos en sus respectivos perfiles en redes sociales.

Las funciones específicas de cada una de las partes que componen esta arquitectura se describen en detalles en las posteriores secciones.

3.2 Sitio de interaccción social: *Facebook*

En 2004, en la Universidad de Harvard se crea *Facebook* como herramienta web a modo de agenda de contactos entre estudiantes compartida en el campus. Desde entonces su crecimiento en número de usuarios activos ha sido enorme,

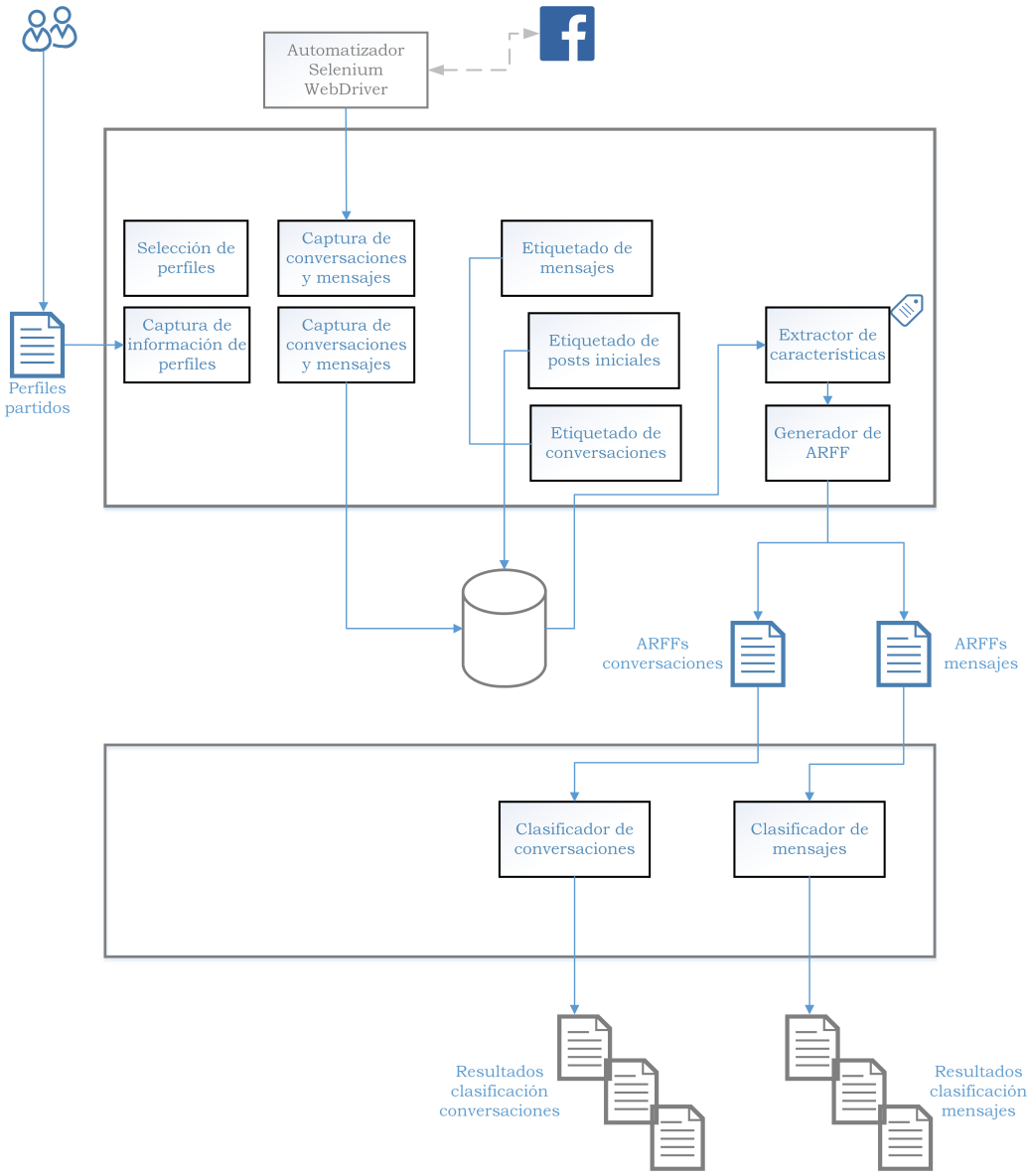


Figura 3.1: Arquitectura general del sistema.

extendiéndose en primer lugar en el mundo universitario y posteriormente en multitud de países. Durante los últimos años ha venido siendo la comunidad *online* con mayor número de usuarios en *Internet*, conectando a personas y colectivos de manera más efectiva, alcanzando en el año 2.014 más de 800 millones de usuarios activos a nivel mundial y convirtiéndose en uno de los principales sitios web en cuanto a volumen de tráfico[1].

En *Facebook* se representan distintos tipos de relaciones entre las personas y colectivos: amistad, pertenencia, y desde 2011 también la relación de parentesco. La relación de amistad es una relación bidireccional, mientras que la de pertenencia y parentesco es unidireccional (que una persona A sea miembro de un grupo B no es recíproco). A partir de las relaciones es como se construye de facto la red social.

3.2.1 Comunicación política entre los miembros de Facebook

Esta plataforma social comenzó a ser utilizada en el ámbito de lo político en los Estados Unidos, desde donde se extendió al resto de sistemas políticos. En ella, el instrumento que se ha utilizado para la comunicación política con mayor frecuencia son los denominados *grupos Facebook*, aunque en los últimos años esta estructura ha sido sustituida por las *Páginas Facebook*, que disponen de una funcionalidad similar a efectos políticos. Tanto a través de los grupos como las páginas, sus miembros o asociados pueden recibir información, invitar a otras personas a unirse, publicar sus propios mensajes, así como enriquecer dicha información incorporando fotos, vídeos y enlaces a otras informaciones y conversaciones, e incluso compartiendo información publicada por terceros [KK09].

Un uso específico de estos *grupos Facebook* son los llamados grupos *partidistas* o grupos vinculados a un determinado partido político, en los que la conversación no es tan abierta, sino que se delimitan los temas y cuestiones a tratar, y en los que con frecuencia se ignoran puntos de vista divergentes o diferentes. Se ha constatado que en términos generales, apenas existen grupos en los que conviven información y temas sobre los que existen opiniones claramente controvertidas [Wes07b]. En este sentido, *Facebook* provee de una infraestructura de comunicación, pero no necesariamente la estructura de una comunidad partidista abierta a cualquier tema de interés que se pudiere suscitar.

Sin ánimo de ser exhaustivo, los partidos políticos han venido utilizando las redes sociales digitales en las siguientes funciones:

[1] Fuente: `http://www.alexa.com/siteinfo/facebook.com`

- **Envío de información política**. En la práctica, los políticos usan *Facebook* para comunicarse con los miembros de su comunidad partidista, a los que se dirigen para transmitirles información con carácter estratégico o táctico. Al comienzo, esta comunicación era principalmente unidireccional, y en cierta medida continúa siendo así en las redes sociales *online*; estos mensajes pueden no ser atendidos y en consecuencia no producir los efectos deseados.

  Sin embargo, no resulta trivial afirmar que una determinada red social consigue forzar a los políticos a leer, reconocer y reaccionar a las respuestas de los electores. La respuesta que los partidos políticos y los candidatos ofrecen a las demandas y opiniones expresadas por la opinión pública ha planteado siempre un desafío para los primeros. Y es precisamente esa una de las tareas claves para que se establezca una verdadera comunicación bidireccional.

- **Cohesionar la comunidad alrededor del partido**. En tanto que red social, *Facebook* sirve de algún modo para aglutinar a los miembros de una comunidad y les proporciona un medio para compartir información a través de una simple red. Por tanto, destaca la función que esta red social desempeña para que los miembros se conecten y se organicen. Y al mismo tiempo, permite a los actores políticos un medio eficaz de llegar a los electores y votantes.

- **Exposición pública del candidato**. La coexistencia de dos esferas de acción política, una de ellas física y otra en línea, ha supuesto un cambio notable en la forma de comunicación de los partidos políticos. En la esfera digital, los candidatos incluyen sus biografías extendidas y resúmenes en sus páginas *Facebook*, información de contacto. Publican álbumes con fotografías y convocan a eventos de campaña. Así, también desarrollan una agenda de publicación de mensajes o *posts* dentro de plataformas de servicio blog y reciben reacciones de respuesta (del inglés *feedback*) a través de esas mismas plataformas. Sin embargo, muchos de los perfiles de candidato dejan de estar activos después de la campaña y otros no mantienen contacto con los electores después de la misma, cuando la mayoría prefieren usar el sitio web oficial para dicha comunicación.

- **Comunicación directa con los candidatos**. Otra capacidad de *Facebook* es la de permitir a los miembros de una comunidad poder enviar mensajes directamente a los candidatos. De este modo, la comunicación e interacción políticas se simplifican y se hacen más directas.

3.3 Adquisición y modelos de datos

Una vez establecido el sitio web de interacción social, pasamos a describir el proceso de toma de datos realizado para la investigación. Este proceso inicial incluye dos tareas; una primera orientada a la recogida automatizada y otra posterior de filtrado realizado de manera manual.

La fase automatizada se ha llevado a cabo a través de un aplicativo software construida explícitamente para tal propósito. Esta solución software se ha desarrollado con tecnología Microsoft .NET y haciendo uso de las plataforma de automatización web *Selenium* en su versión 2.44 y *WebDriver* 2.45 [1]. Toda la información recogida a lo largo del proceso se ha almacenado en una base de datos relacional SQL Server para su posterior tratamiento.

Este proceso de descarga automática de la información se ha utilizado para la recolección de información de los distintos perfiles de los partidos políticos, así como las conversaciones públicas y la información pública de los perfiles de los ciudadanos.

Para la selección del período de captura de la información, se he han llevado sucesivas tomas de muestras, con el propósito de encontrar el más propicio. De este modo, durante las elecciones de 2011 se realizaron muestreos de información, orientados a dos objetivos principales. En primer lugar, trataba de obtener información para poder representar la estructura relacional de identidades digitales que se construye alrededor de los muros de los partidos políticos en las redes sociales. Esto es, las redes más próximas y más alejadas de los partidos políticos, y que marcan distintos niveles de influencia social basada en la distancia al foco origen.

En segundo lugar, se recogió información sobre las conversaciones provenientes de todos los perfiles que los dos partidos principales, Partido Popular (PP) y Partido Socialista Obrero Español (PSOE), tienen distribuidos orgánica y geográficamente por toda España, y a través de los cuales se estaba realizando durante la campaña electoral de aquel año. Dada la heterogeneidad y fragmentación de los mensajes, no se alcanzaron resultados significativos respecto a las respectivas estrategias de comunicación de los partidos a través de sus perfiles en redes sociales.

Durante las elecciones europeas de mayo de 2014, la emergencia de un nuevo partido político, *Podemos*, supuso una alteración del sistema bipartidista de efectos indeterminados. Unido a esta circunstancia se da el hecho de que históricamente las elecciones europeas no son representativas respecto a las dinámicas

[1] `http://www.seleniumhq.org`

en elecciones autonómicas o elecciones generales en España.

Finalmente, la última recogida de datos se realiza con motivo de las elecciones autonómicas y locales de 2015. Los datos obtenidos son las conversaciones políticas que surgen en los muros digitales de dichos perfiles sociales durante el período de pre-campaña y campaña electoral que va desde el 1 al 22 de mayo de 2015. En este caso, en lugar de recoger toda la información de todos los perfiles digitales en la red social *Facebook* de los que disponen los dos principales partidos políticos en España, se diseñó una estrategia distinta. En primer lugar, se obtuvo la información de tres partidos políticos: *PP*, *PSOE*, y *Podemos*. Además de esto, En segundo lugar, se recogió la información publicada en los perfiles de los propios candidatos en la red social *Facebook*, que reportaba información muy similar a la de los perfiles de sus respectivos partidos políticos. En total se han recogido 6.845 mensajes contenidos en 79 conversaciones en los muros digitales de los partidos políticos señalados.

Cabe señalar que los datos relativos a la identidad personal de los perfiles en *Facebook* de los ciudadanos que han participado en las conversaciones *online*, su información se ha tratado de manera anónima, ya que su identidad nominal no es objeto de estudio. Es por este motivo que a cada identidad personal nominativa le ha sido asignada un identificador numérico único que representa de manera anónima a cada perfil de ciudadano que ha participado en cualesquiera conversación.

La validación de la autenticidad de los perfiles políticos se ha realizado accediendo directamente a los perfiles *Facebook* de las identidades digitales y comprobando la veracidad de los datos de información general reportados por ellos mismos.

A continuación se describen los modelos de datos que se han utilizado durante la investigación. En primer lugar pasaremos a detallar el conjunto de datos de los mensajes para, una vez descritos, explicar el modelo del conjunto de datos de las conversaciones políticas que tienen lugar en los muros *online*.

3.3.1 **Conjunto de datos de mensajes políticos**

Antes de detallar la definición y estructura de un mensaje, resulta pertinente señalar que el espacio en el que tienen lugar todas las interacciones es el *muro digital*, o simplemente *muro* del perfil de un partido político, tal y como se puede apreciar en la figura 3.2.

Los mensajes son publicados en los muros de los partidos políticos, o bien surgen como respuesta a un *post* o mensaje inicial que el titular del muro escribe



Figura 3.2: Ejemplo de *muro* digital.

en su muro. El mensaje es el elemento mínimo de una conversación.

Un ejemplo de mensaje político es el que se muestra en la figura 3.3. Como se puede observar, el perfil de un político en *Facebook* se estructura visualmente según paneles. Gran parte de estos contenidos varían con el tiempo, y muchos pueden personalizarse individualmente por el propietario del perfil. En el panel superior aparece la imagen del perfil y su nombre. Además, incluye enlaces a distintas secciones o prestaciones (la biografía, la información general, los álbumes de fotos y otras secciones adicionales).

En el panel izquierdo se muestra información acerca de la notoriedad del perfil, tal como el número de seguidores o personas interesadas, o las personas que vinculan al perfil político con el usuario en curso. Junto a estos datos, aparece información gráfica relevante, tal como fotos o vídeos.

El panel central (en la figura, a la derecha) se presenta el muro del perfil político. Este muro representa una suerte de tablero digital en el que se van publicando sucesivos mensajes que contienen diferentes tipos y formatos de contenidos.

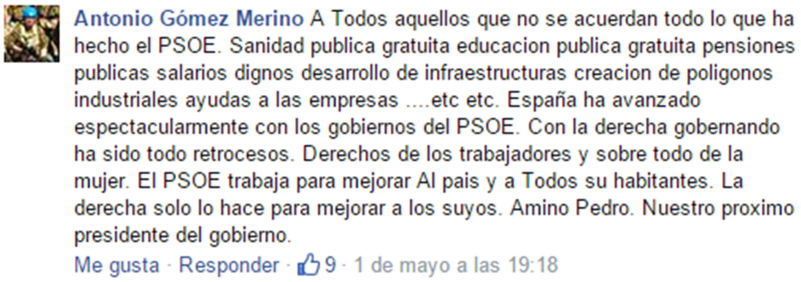


Figura 3.3: Estructura de un mensaje dentro de una conversación política.

Cada uno de los mensajes cuenta con una estructura y características que se enumeran a continuación:

- **Fecha**. Momento en el que se publica el mensaje.
- **Imagen**. Avatar o identidad gráfica del autor. Es frecuente que eta imagen contenga una fotografía de la persona a la que representa, pero esta circunstancia no es obligatoria.
- **Autor**. Identidad de la persona que publica el mensaje, esto es, con nombre y apellidos. Sin embargo, en los últimos tiempos están surgiendo nuevas formas de autodenominarse en las redes sociales. En todo caso, con el objeto de garantizar el carácter anónimo, esta información se ha tratado de manera absolutamente transparente, asignando a cada autor un identificador numérico.
- **Identificador del mensaje**. Esta información, que no queda visualizada gráficamente en el sitio social, es un número que representa de manera biunívoca a un mensaje dentro de una conversación.
- **Orden de secuencia**. El orden correlativo del mensaje dentro de una conversación iniciada a partir de un post. Permite reconocer y reconstruir el hilo temporal de una conversación.
- **Cuerpo del mensaje**. Es a través del cuerpo del mensaje donde el ciudadano se expresa con respecto a cualquier tema. Las condiciones de esa expresión quedan limitados a las políticas que *Facebook* establece a tal

efecto, y que fomentan un comportamiento respetuoso y cívico en las manifestaciones que se viertan a través de dicha red social [1]. El cuerpo del mensaje, además de texto, puede contener información no textual, tal como enlaces, abreviaturas y otras expresiones.

- **Grado de aprobación**. Indica el número de veces que se ha indicado aceptación del mensaje a través de la funcionalidad "*Me gusta*".

No todas las características del conjunto de datos de mensajes se han incluido como dato de entrada en la tarea de análisis, sino solamente aquellas cuya información aportada es relevante y no introduce distorsión en el conjunto final de datos.

3.3.2 Conjunto de datos de conversaciones políticas

El conjunto de datos de las conversaciones políticas contiene el total de los mensajes que se inician con un mensaje de un perfil de partido y que pueden generar respuestas de identidades que están conectadas a ellos. Estas conversaciones forman parte y contribuyen al debate político y por tanto, también del debate democrático.

La Figura 3.4 muestra el fragmento de una conversación iniciada por un perfil de partido político.

- **Autor**. Identifica el perfil digital que inicia la conversación en su muro.
- **Fecha**. Momento en el que se origina la conversación.
- **Post o mensaje inicial**. Representa el primer mensaje de la conversación. Generalmente se trata de un mensaje escrito por el propio perfil, al que responden ciudadanos. La figura 3.5 muestra la representación de la cabecera una conversación iniciada por un perfil de partido político.
- **Grado de aprobación**. Recoge el número de personas distintas que han mostrado agrado, acuerdo o conformidad con el mensaje, haciéndolo explícito a través del botón "*Me gusta*".
- **Identificador conversación**. Es un identificador numérico que señala biunívocamente cada una de las conversaciones que tienen lugar. Del mismo modo que en el caso de los mensajes, el identificador no se muestra gráficamente.

[1] `https://www.facebook.com/communitystandards`



Figura 3.4: Estructura de una conversación en *Facebook*.

- **Grado de compartición**. Es el número de veces que los usuarios han compartido la conversación entres sus personas de contacto. Equivale al número de personas que han compartido el contenido de la conversación.

- **Lista de mensajes de respuesta**. Es el conjunto de mensajes que los usua-



Figura 3.5: Estructura de un *post* o mensaje inicial en una conversación.

rios han emitido como respuesta al mensaje inicial. La estructura de un mensaje se ha descrito anteriormente, y se muestra en la Figura 3.3.

De manera análoga al tratamiento del conjunto de datos de los mensajes, no todas las características del conjunto de datos de las conversaciones han sido incorporadas para su análisis, sino solamente aquellas cuya información aportada es relevante y no introduce ruido informacional en el conjunto final de datos.

3.4 Etiquetado de los conjuntos de datos

Tras la recolección automática de datos, procede el etiquetado manual de los mismos, dentro del proceso definido por las técnicas de aprendizaje supervisado. Para ello, según se ha indicado en la sección anterior recogida toda la información sobre los contenidos vertidos en los perfiles sociales, se han filtrado mensajes publicados que no están orientados hacia una comunicación política explícita, sino a comentar la información de carácter general, mayoritariamente irrelevante para el análisis politológico.

Para cada uno de los conjuntos de datos se han definido distintas clasificaciones que se describen a continuación:

3.4.1 Conjunto de mensajes en muros de partidos políticos

La globalidad de los mensajes publicados en muros de los perfiles de partidos políticos se han clasificado atendiendo a distintos criterios.

A fin de clarificar las posteriores explicaciones, la Tabla 3.1 contiene los distintos tipos de clasificación que se llevan a cabo sobre los mensajes que se explican posteriormente.

Tabla 3.1: Dimensiones de los mensajes y sus clasificaciones.

Dimensiones	Clasificaciones
A nivel funcional	*Función informativa* *Función movilizadora*
A nivel deliberativo	*Grado de contribución* *Grado de debate político* *Referencialidad de los contenidos*
Análisis argumentativo	*Nivel de personalización* *Nivel de metacomunicación* *Nivel de campaña negativa*
Estrategia comunicacional	*Marco primario* *Marco secundario*

3.4.1.1 Clasificaciones del análisis funcional

Este tipo de clasificación recoge las funciones principales que desempeña un determinado mensaje o contenido comunicacional, que son las siguientes:

1. **Según su función informativa**. Si un mensaje tiene un carácter eminentemente informativo, a su vez puede reportar respecto a distintos tipos de información. En consecuencia, un mensaje político respecto a su función informativa puede ser de alguno de estos tipos:

   (a) *Informativo con enlace a contenidos del propio partido*. El mensaje informa referenciando al propio partido político. Véase la Figura 3.6.

   (b) *Informativo con enlace a contenidos de terceros*. El mensaje informa, incorporando referencias de terceros (medios de comunicación, instituciones, ciudadanos u otros). Véase la Figura 3.7.



Figura 3.6: Ejemplo de mensaje informativo con referencia a contenidos propios.



Figura 3.7: Ejemplo de mensaje informativo con referencia a contenidos de terceros.

2. **Según su función movilizadora**. El mensaje puede incitar a la movilización política desde distintos niveles. Dependiendo del objetivo que pretende la movilización del mensaje, éste puede ser:

   (a) *Movilizador hacia el voto*. El mensaje mueve a que se vote, indepen-

dientemente del sentido del voto. Véase la Figura 3.8.

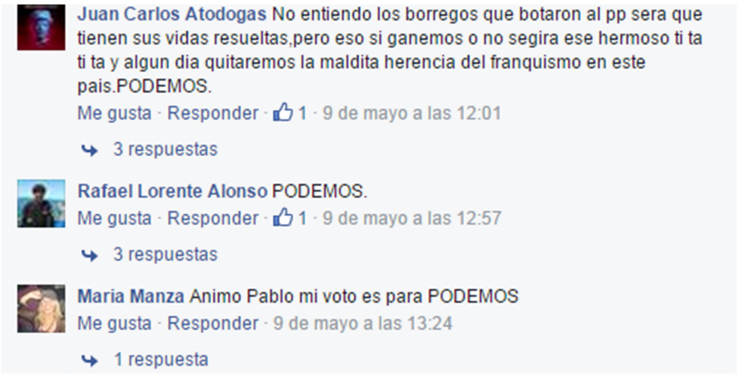


Figura 3.8: Ejemplo de mensaje de tipo movilizador hacia el voto.

(b) *Movilizador hacia otra acción política*. El mensaje trata de promover una acción política distinta al voto, ya sean formas convencionales o no convencionales de participación política (ej. participación en eventos políticos, manifestaciones, etc.). Véase la Figura 3.9.

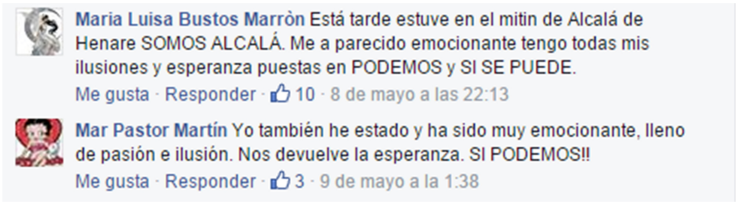


Figura 3.9: Ejemplo de mensaje de tipo movilización hacia otra acción política.

3.4.1.2 **Clasificaciones del nivel deliberativo**

Desde el punto de vista de la deliberación democrática, se han establecido las siguientes clasificaciones de los mensajes:

1. **Según el grado contribución informativa**. Indica si el mensaje aporta o no información, valoraciones o contenidos a la discusión política

   (a) *Contribución nula*. Este tipo de mensajes no aporta ni información ni opinión a la discusión política.

   (b) *Contribución neta*. El mensaje sí aporta información, opiniones o contenidos a la conversación.

1. **Según el grado de debate político**. El grado de una discusión o conversación política se construye en base al carácter de los mensajes de dicha discusión. Según esta clasificación, un mensaje político pueden contribuir a la discusión política en uno de los siguientes niveles:

   (a) *Irrelevante*. Este tipo de mensajes no aporta ni información ni opinión a la discusión política. Véase la Figura 3.10.

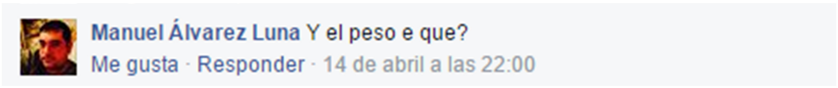


Figura 3.10: Ejemplo de mensaje de debate de tipo irrelevante.

   (b) *Reafirmativo*. Refuerza el argumento o idea expresado en el mensaje inicial de la conversación. Véase la Figura 3.11.

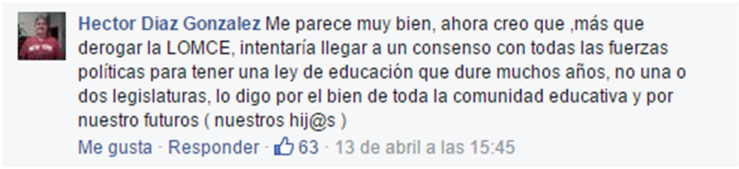


Figura 3.11: Ejemplo de mensaje de debate de tipo reafirmativo.

   (c) *Controvertido*. Pone en cuestión la tesis o idea manifestada en el mensaje inicial. Véase destacado en azul en la Figura 3.12.



Figura 3.12: Ejemplo de mensaje de debate de tipo controvertido.

(d) *Polémico*. Pretende polemizar sin necesidad de debatir o llegar a una conclusión dentro del debate, o realizando afirmaciones que suponen una clara confrontación con respecto al mensaje inicial. Véase la Figura 3.13.

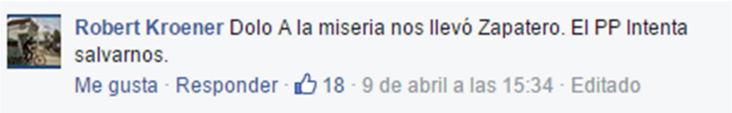


Figura 3.13: Ejemplo de mensaje de debate de tipo polémico.

(e) *Insulto*. El comentario supone una ofensa o falta de respeto, bien al mensaje, a su emisor, a las personas que participan en la discusión o a terceros. Véase la Figura 3.14.

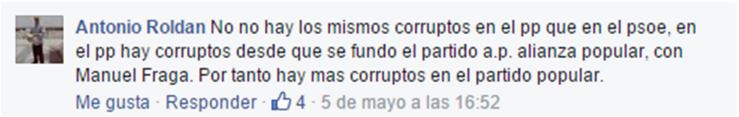


Figura 3.14: Ejemplo de mensaje de debate de tipo ofensivo o insulto.

2. **Según la referencialidad de los contenidos**. El mensaje puede incorporar contenidos adicionales, que pueden haber sido generados y publicados por el propio partido político, por un medio de comunicación o bien por un tercero distinto a cualquiera de ellos.

(a) *Sin referencia*. El mensaje no contiene referencia alguna a otro contenido tercero. Véase la Figura 3.15.

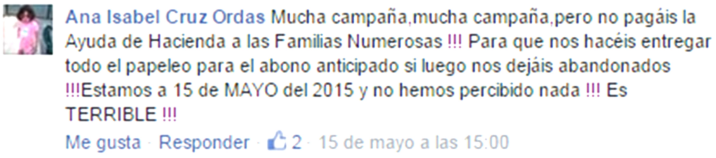


Figura 3.15: Ejemplo de mensaje sin referencias externas.

(b) *Referencia a otros*. El contenido referenciado ha sido publicado en por otro distinto a un medio o al propio partido. Véase la Figura 3.16.

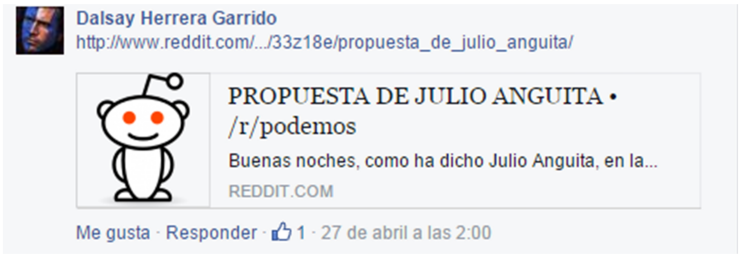


Figura 3.16: Ejemplo de mensaje que referencia a contenidos de terceros.

(c) *Autorreferencial*. Los contenidos son de la web del partido o la del candidato. Véase la Figura 3.17.



Figura 3.17: Ejemplo de mensaje de tipo autorreferencial.

(d) *Referencia a un medio de comunicación*. El contenido al que se hace referencia ha sido publicado por un medio de comunicación. Véase la Figura 3.18.

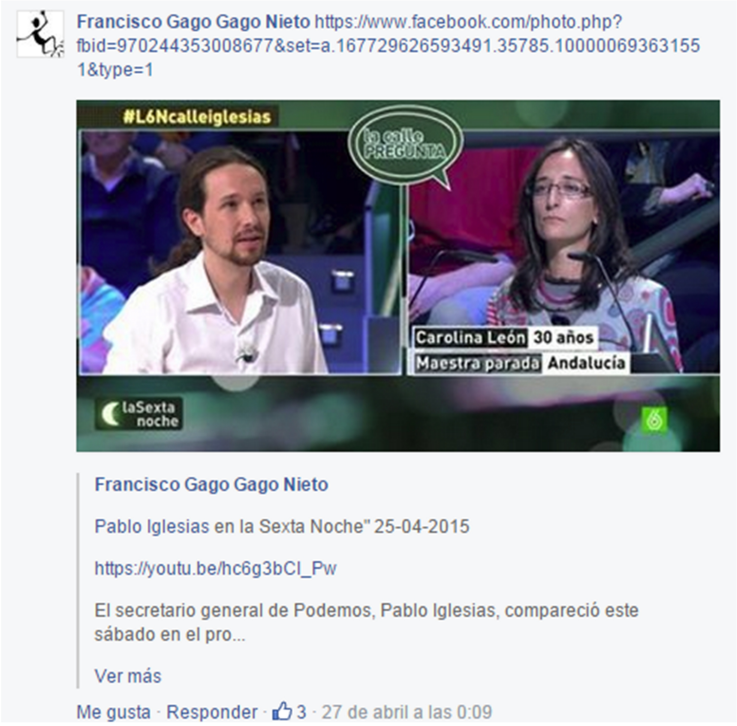


Figura 3.18: Ejemplo de mensaje que referencia a contenidos de un medio de comunicación.

3.4.1.3 **Clasificaciones del análisis argumentativo**

Desde el análisis argumentativo, el mensaje puede operar a varios niveles de comunicación: a nivel de personalización, a nivel de meta-comunicación, o a nivel de campaña negativa.

(a) **Nivel de personalización**. Esta clasificación describe sobre quién se centra o a quién se refiere el contenido del mensaje.

i. *Mención al propio candidato*. El mensaje realiza menciones al propio candidato. Véase la Figura 3.19.



Figura 3.19: Ejemplo de mensaje de personalización hacia el propio candidato.

ii. *Mención al partido propio*. El mensaje hace referencias al partido al que pertenece el perfil. Véase la Figura 3.20.

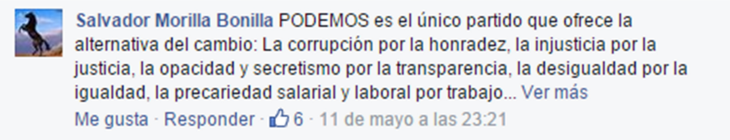


Figura 3.20: Ejemplo de mensaje de personalización hacia el propio partido.

iii. *Mención a otros candidatos*. El mensaje nombra o hace referencia a otros candidatos. Véase la Figura 3.21.



Figura 3.21: Ejemplo de mensaje de personalización hacia el otro candidato.

(b) **Nivel de meta-comunicación**

i. *Eslóganes electorales*. El mensaje utiliza eslóganes electorales. Véase la Figura 3.22.

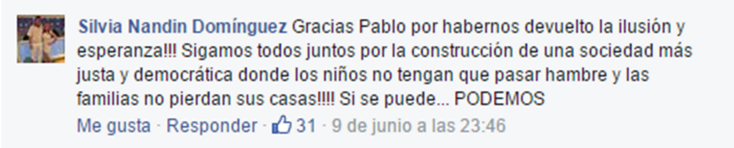


Figura 3.22: Ejemplo de mensaje de tipo meta-comunicación con uso de eslóganes.

ii. *Propuestas políticas*. El mensaje incluye proposiciones a realizar a los candidatos, al partido o a una institución. Véase la Figura 3.23.

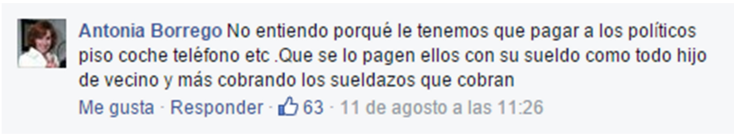


Figura 3.23: Ejemplo de mensaje de tipo meta-comunicación con propuestas de acción política.

iii. *Cobertura de la campaña electoral*. El mensaje hace referencia a la campaña electoral. Véase la Figura 3.24.

(c) **Nivel de campaña negativa**. Los mensajes muestran una campaña en clave negativa, bien insistiendo en aspectos negativos o pesimistas, bien marcando los errores, defectos o carencias del adversario.

i. *Menciones a los partidos adversarios*. La mención negativa se hace a los partidos oponentes. Véase la Figura 3.25.

ii. *Menciones a los candidatos adversarios*. La orientación negativa del mensaje es respecto al candidato o candidatos oponentes. Véase la Figura 3.26.



Figura 3.24: Ejemplo de mensaje de meta-comunicación con cobertura de campaña.

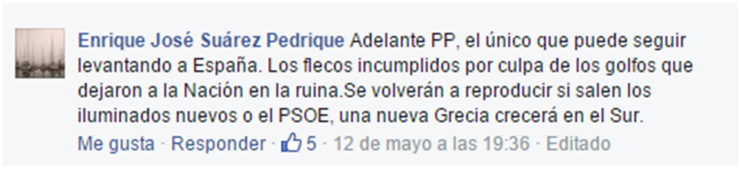


Figura 3.25: Ejemplo de mensaje de tipo campaña negativa con mención a partidos adversarios.

3.4.1.4 Clasificaciones de la estrategia comunicativa

Esta clasificación atiende a la presencia de marcos comunicacionales en los mensajes. Según ello, los mensajes pueden tener activos distintos tipos de encuadre. El marco puede operar en dos niveles de influencia, denominados primario y secundario. Tanto en un nivel u otro, los marcos pueden ser de los tipos siguientes:

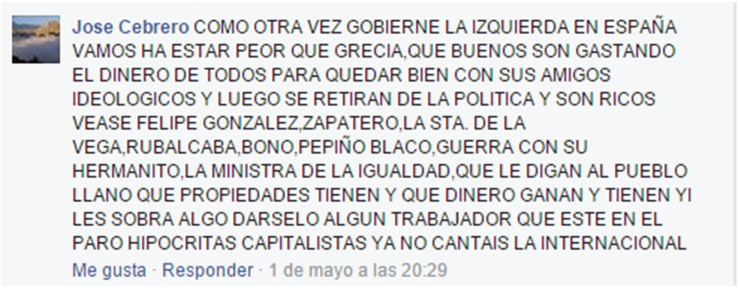


Figura 3.26: Ejemplo de mensaje de tipo campaña negativa con mención a candidatos adversarios.

1. **Marco Primario**

   (a) *Marco de conflicto*. El mensaje presenta una información evidenciando o enfatizando la tensión o conflicto entre individuos, grupos o instituciones. Véase la Figura 3.27.

   (b) *Marco de interés humano*. El mensaje plantea un evento, tema o problemática como la historia de un individuo o desde un enfoque eminentemente emocional, como muestra la Figura 3.28.

   (c) *Marco de consecuencias económicas*. El mensasje plantea los temas en relación con las consecuencias económicas que conlleva o bien desde el punto de vista meramente racional o cuantitativo. Véase la Figura 3.29.

   (d) *Marco de responsabilidad*. El marco principal es el de atribución de responsabilidad desde el punto de vista moral o religioso. Véase la Figura 3.30.



Figura 3.27: Ejemplo de mensaje con marco de tipo conflicto.



Figura 3.28: Ejemplo de mensaje con marco de tipo interés humano.



Figura 3.29: Ejemplo de mensaje con marco de tipo consecuencias económicas.

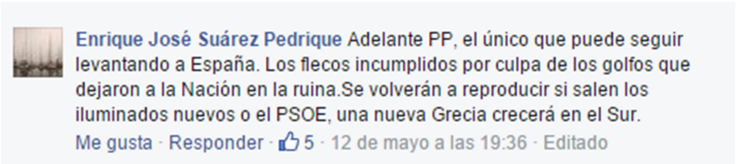


Figura 3.30: Ejemplo de mensaje con marco de tipo responsabilidad.

En caso de que el marco activo sea el de responsabilidad, ésta se puede atribuir de manera distinta y a distintos roles.

(a) *Oponente inmoral*. Cuando el mensaje proyecta al adversario político como inmoral o ilegítimo, presentando un fraude demostrado del oponente, identificando al contrincante como enemigo del pueblo o bien al oponente como un líder fallido. Véase la Figura 3.31.

(b) *Candidato populista*. En el contenido del mensaje el candidato o el

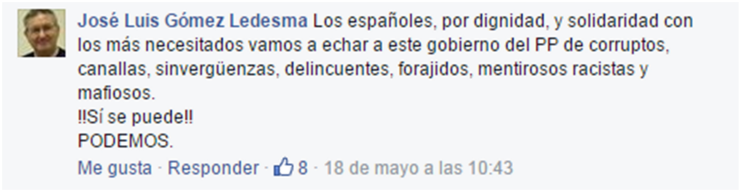


Figura 3.31: Ejemplo de mensaje con marco moral de tipo oponente inmoral.

partido es confiable y honesto; el candidato o líder es independiente y justo, o bien comparte los valores del país o territorio. El ejemplo de la Figura 3.32 muestra un caso de este tipo de mensajes.

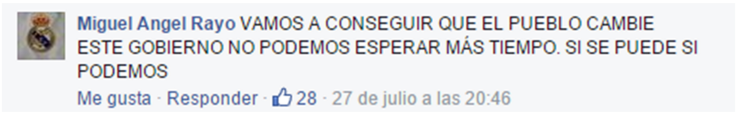


Figura 3.32: Ejemplo de mensaje con marco moral de tipo candidato populista.

(c) *Candidato como héroe*. El candidato aparece como una persona que ha sobrevivido o superará unas altas exigencias, retos de difícil cumplimiento, o bien que se ha sobrepuesto a dificultades personales a lo largo de su experiencia de vida. Se trata de un líder que pelea contra los enemigos o adversarios, con el fin de obtener un objetivo mayor, o para proteger al pueblo de terceros. En la Figura 3.33 se muestra un ejemplo.

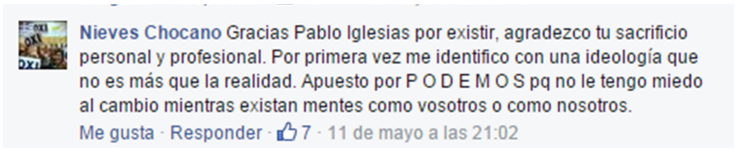


Figura 3.33: Ejemplo de mensaje con marco de tipo candidato héroe.

(d) *Candidato como estratega*. En el mensaje, se presenta al candidato como una persona que hace cálculos electoralistas, de gobierno o de partido. Se propone a un líder o candidato progresista y/o pragmático, tal como muestra la Figura 3.34.

(e) *Candidato como víctima*. Cuando el candidato es objeto de ataques políticos, profesionales o personales por parte de sus adversarios políticos. Véase la Figura 3.35.

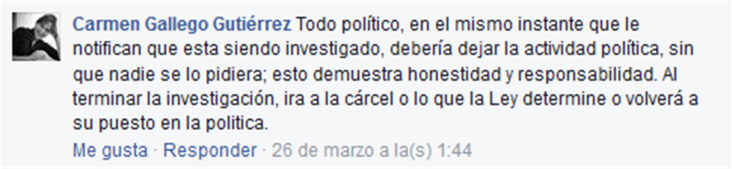


Figura 3.34: Ejemplo de mensaje con marco moral de tipo candidato estratega.

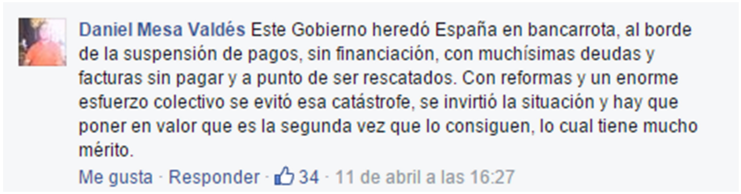


Figura 3.35: Ejemplo de mensaje con marco de tipo candidato víctima.

2. **Marco secundario**. Cuando un marco comunicacional se activa puede hacerlo con dos intensidades o a dos niveles. El marco secundario represente el nivel de menos influencia de los dos. Su análisis se realiza utilizando los mismos criterios que el marco primario, explicado anteriormente. El marco primario y secundario se pueden activar simultáneamente y pueden ser del mismo tipo, para lo cual, los asuntos que sean objeto de enmarcado deben ser necesariamente diferentes.

3.4.2 Conjunto de conversaciones en muros de partidos políticos

Las discusiones que tienen lugar en el muro de un partido político se clasifican a tenor de distintos criterios. A saber, según la función que desempeñan, la pluralidad o heterogeneidad, según el nivel de información, el grado de debate político, la referencialidad o el tipo de argumentación que se desarrolla.

Establecidas las clases, se ha llevado a cabo un proceso de etiquetado manual de cada una de las instancias que representan mensajes en una conversación política en red social. Este trabajo se ha realizado mediante el uso de un aplicativo software construido a tal efecto. En la Figura 3.2 se muestra un ejemplo en el que aparece el resultado del proceso de etiquetado de un mensaje político.

El conjunto de datos de las conversaciones ha sido construido a partir de varios conjuntos de datos. En primer lugar, por cada conversación se ha incluido información general del perfil que inicia la conversación. Además, se incluye el *post* o mensaje inicial emitido por el partido político en su muro, que ha sido

etiquetado individualmente. Seguidamente, cada conversación lleva incluido el conjunto de datos de todos los mensajes que se han generado como respuesta a dicho *post* inicial, y que se ha descrito con anterioridad.

Tabla 3.2: Dimensiones de las conversaciones y sus clasificaciones.

Dimensiones	**Clasificaciones**
A nivel funcional	*Función informativa* *Función movilizadora*
Análisis argumentativo	*Nivel de personalización* *Nivel de metacomunicación* *Nivel de campaña negativa*
Análisis de marcos	*Marco primario* *Marco secundario*

3.4.2.1 Clasificaciones del análisis funcional

1. **Según su función informativa** De manera análoga a los mensajes políticos publicados, las conversaciones pueden desempeñar una función informativa:

   (a) *Informa sin enlaces a terceros*. El propósito general de la conversacion consiste en informar sobre un asunto, más que en hacer referencias permanentes a terceras fuentes.

   (b) *Informa con enlace a contenidos del propio partido*. La conversación emite información a partir de las distintas interacciones o mensajes de los participantes que aportan contenidos del propio partido político del muro.

   (c) *Informa con enlace a contenidos de medios de comunicación*. De igual forma que en el caso anterior, la conversación fluye con aportaciones continuas de contenidos de terceros que son medios de comunicación.

   (d) *Informa con enlace a contenidos de terceros*. Se informa, incorporando referencias de terceros (medios de comunicación, instituciones, ciudadanos u otros).

2. **Según su función de movilización** Las conversaciones pueden cumplir una función eminentemente de movilización de personas y recursos para la acción política. Los dos tipos de movilización son:

(a) *Moviliza con apelación al voto*. La conversación mantiene un tono general de incitación a la movilización, invitación a la participación política o incluso al votar, independientemente del sentido del voto.

(b) *Moviliza con apelación a la movilización*. El tipo de movilización al que invita o promueve no es electoral, sino que trata de mover a cualquier otro tipo de acción política diferente al voto.

3.4.2.2 **Clasificaciones del análisis argumentativo**

1. **Nivel de personalización**

   (a) *Menciones al candidato propio*. La conversación contiene un alto nivel de menciones al candidato o máximo responsable político del sitio social *online* en cuyo muro se publica el mensaje.

   (b) *Menciones a otros candidatos*. A lo largo de la conversación las alusiones a un candidato distinto al del partido en cuyo muro aparece publicado el mensaje.

   (c) *Menciones al partido*. Durante el transcurso de la conversación se hace mención con frecuencia al partido político en cuyo muro digital se publica el mensaje.

2. **Nivel de meta-comunicación**

   (a) *Cobertura de la campaña electoral*. El mensaje hace mención a aspectos generales de campaña, tales como la celebración de eventos, mítines, las fechas o momentos claves de la campaña o al propio hecho de estar en campaña.

   (b) *Eslóganes electorales*. El mensaje incluye lemas, eslóganes o arengas diseñadas por los partidos durante la campaña de la competición electoral.

   (c) *Propuestas de acción política*. El mensaje incluye recomendaciones, propuestas o sugerencias para que sean llevadas a cabo, bien por algún candidato, por un partido o por una institución política, tal como un gobierno.

3. **Nivel de campaña negativa**

   (a) *Menciones al partido adversario*. Este tipo de referencias son alusivas al principal partido adversario (entre Partido Popular, Partido Socialista Obrero Español o Podemos) o a cualquiera de los que compiten electoralmente.

(b) *Menciones a los candidatos adversarios*. La campaña negativa se centra, no en el partido político opuesto, sino el candidatos o máximos responsables de dichos partidos.

3.4.2.3 **Clasificaciones del análisis de marcos**

Esta clasificación atiende a la presencia de marcos comunicacionales en los mensajes. Según esta clasificación, los mensajes pueden tener activos distintos tipos de encuadre. El marco puede operar en dos niveles de influencia o efecto. Tanto en un nivel u otro, los marcos pueden ser de los mismos tipos.

1. *Marco de conflicto*. El conjunto de los mensajes generados tras el mensaje inicial en la conversación evidencian o enfatizan la tensión o conflicto entre individuos, grupos o instituciones.

2. *Marco de interés humano*. La discusión política transcurre en términos de un suceso, tema o problemática que es la historia de un individuo o desde un enfoque eminentemente emocional.

3. *Marco de responsabilidad*. El marco principal de la conversación política es el de atribución de responsabilidad desde el punto de vista moral o religioso.

4. *Marco de consecuencias económicas*. El flujo de los mensajes de la conversación responden planteando las consecuencias económicas de un determinado o varios asuntos.

En este punto cabe recordar que, al igual que sucede para los mensajes políticos, en el caso de que el marco activo (ya sea a nivel primario o secundario) sea el de responsabilidad, ésta se puede atribuir de manera distinta y a distintos roles, que son los siguientes:

1. *Oponente inmoral o inmoralidad del oponente*. Cuando el hilo de la conversación proyecta al adversario político como inmoral o ilegítimo, presentando un fraude demostrado del oponente, identificando al contrincante como enemigo del pueblo o bien al oponente como un líder fallido.

2. *Candidato populista*. A lo largo de la conversación, el candidato o el partido es confiable y honesto. Se le atribuyen valores como justicia e independencia, al mismo tiempo que el candidato es reconocido por compartir los valores del país o territorio a los que representa o dice representar.

3. *Candidato como héroe*. El candidato aparece como una persona que ha sobrevivido o superará unas altas exigencias, retos de difícil cumplimiento, o bien que se ha sobrepuesto a dificultades personales a lo largo de su experiencia de vida. Se plantea en términos de batalla contra los enemigos, como protector del pueblo, y se le atribuye firmeza y trabajo.

4. *Candidato como estratega*. En las distintas interacciones o mensajes de la discusión política, se presenta al candidato como una persona que hace cálculos electoralistas, de gobierno o de partido. Por tanto, se ofrece la figura de un candidato pragmático y progresista u orientado hacia el futuro.

5. *Candidato como víctima*. Cuando el candidato es objeto de ataques políticos, profesionales o personales por parte de sus adversarios políticos. Se muestra a un candidato acusado erróneamente, cuya reputación ha sido dañada, e incluso atacado por sus propios compañeros.

3.4.3 Discusión de los resultados del etiquetado

Consideremos ahora los resultados del proceso de etiquetado de los mensajes y las conversaciones políticas que tienen lugar en los muros digitales de los partidos. Se presentan dichos resultados ordenados para cada uno de los conjuntos de datos sobre los que se han realizado el etiquetado.

Recordemos que el proceso de etiquetado se ha realizado de manera manual utilizando un aplicativo software desarrollado a tal efecto, que permite el registro de las etiquetas en la base de datos SQL Server referida al comienzo del capítulo.

3.4.3.1 Mensajes políticos

El proceso de etiquetado de mensajes en la red social según cada uno de los clasificadores ha producido distintas distribuciones de cada una de sus respectivas clases. Estas distribuciones se comentan a continuación.

En primer lugar, la clasificación por función informativa, los mensajes se distribuyen de manera descompensada entre las distintos tipos, predominando mayoritariamente aquellos mensajes que no incluyen ningún tipo de referencia, ya fueren propias o hacia terceros. En la Tabla 3.3 se recogen estos datos.

Los mensajes publicados con intención movilizadora se distribuyen en torno a las distintas clases en la forma que se describe en la Tabla 3.4. Estos datos

Tabla 3.3: Distribución de mensajes con respecto a la función informativa.

Clase	Número de mensajes	(%)
Función informativa sin enlaces	6.411	93,66
Función informativa con enlaces propios	29	0,42
Función informativa con enlaces a terceros	405	5,92
Total	6.845	100

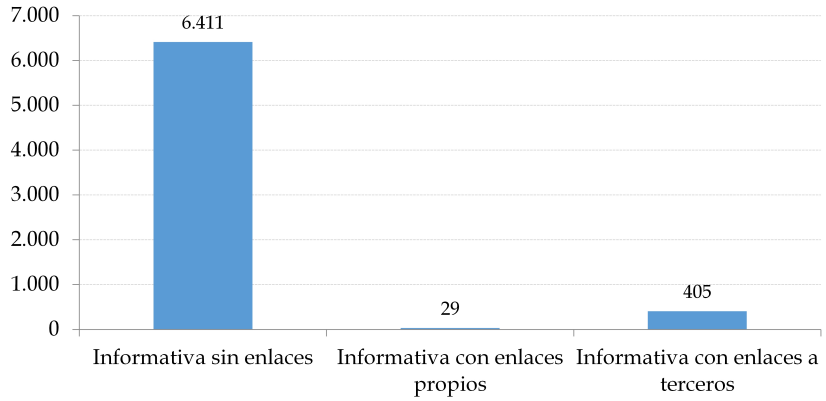


Figura 3.36: Histograma con la distribución de mensajes respecto a la función informativa.

señalan que tres cuartas partes de los mensajes publicados realizan aportaciones significativas al la discusión política de la que forman parte. Para mayor claridad, véase la Figura 3.37.

Tabla 3.4: Distribución de mensajes con respecto a la función movilizadora.

Clase	Número de mensajes	%
N/A	6049	88,37
Movilizacíon hacia el voto	408	5,96
Movilizacíon hacia otros tipos de acción	388	5,67
Total	6.845	100

La distribución de las clases que muestran el grado de contribución de un determinado mensaje político se muestran en la siguiente Tabla 3.5. Estos datos señalan que tres cuartas partes de los mensajes publicados realizan aportaciones significativas a la discusión política de la que forman parte. Para mayor claridad, véase la Figura 3.38.

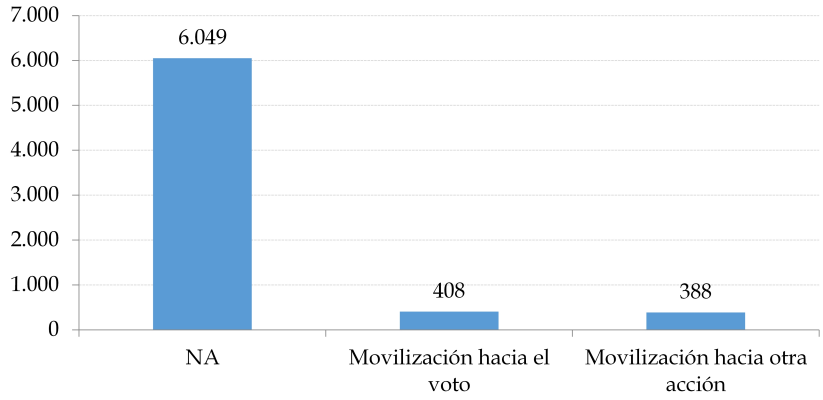


Figura 3.37: Histograma con la distribución de mensajes respecto a la función de movilización.

Tabla 3.5: Distribución de mensajes con respecto al grado de contribución.

Clase	Número de mensajes	%
Contribución nula	1.755	25,64
Contribución neta	5.090	74,36
Total	6.845	100

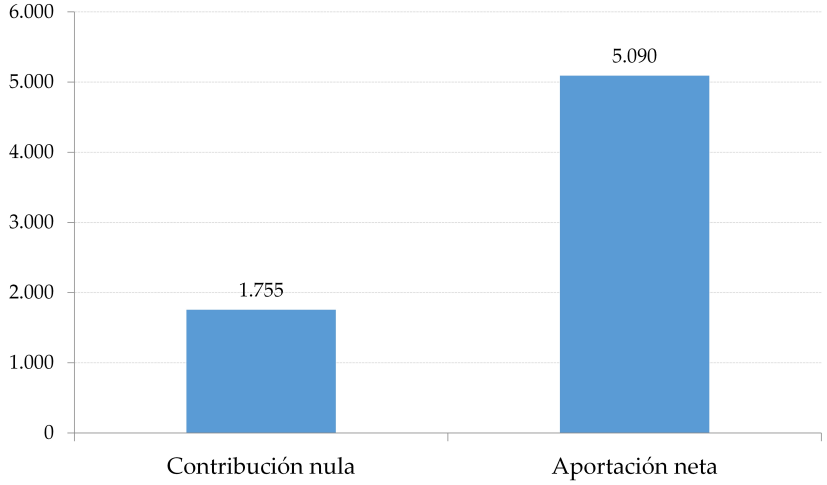


Figura 3.38: Histograma con la distribución de mensajes respecto a la función informativa.

Por su parte, los mensajes que muestran el grado de debate dentro de una discusión política se distribuyen tal como muestra la Tabla 3.6. Estos datos se-

ñalan que tres cuartas partes de ellos realizan aportaciones significativas a la conversación política de la que forman parte. Para mayor claridad, véase la Figura 3.39.

Tabla 3.6: Distribución de mensajes con respecto al grado de debate político.

Clase	Número de mensajes	%
Irrelevante	990	14,46
Reafirmativo	4073	59,50
Controvertido	830	12,13
Polémico	815	11,91
Insulto u ofensa	137	2,00
Total	6.845	100

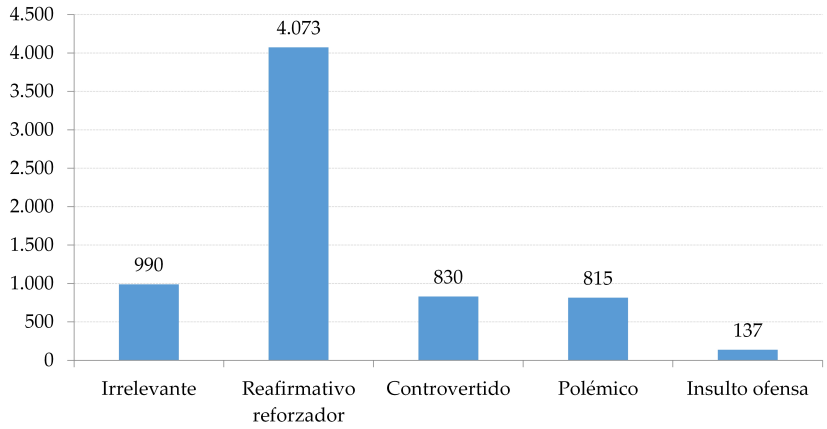


Figura 3.39: Histograma con la distribución de mensajes respecto al grado de debate político.

En lo que se refiere al análisis de la meta-comunicación, la distribución de las mensajes en torno a la clasificación del nivel de meta-comunicación señalan que tan solo una quinta parte de los mensajes presenta un nivel de meta-comunicación de campaña significativo en términos de la cobertura del proceso electoral, mensajes con propuestas de acción política o bien el uso de eslóganes electorales. La Tabla 3.7 muestra la distribución de los resultados. En la figura 3.40 se representa gráficamente esta distribución.

La distribución de frecuencias de cada una de las clases asociadas a la referencialidad de los contenidos se muestra en la Tabla 3.8. Estos datos señalan que la mayoría de los mensajes publicados no ofrecen vinculación a otros conte-

Tabla 3.7: Distribución de mensajes según el nivel de meta-comunicación.

Clase	Número de mensajes	%
N/A	5300	77,43
Cobertura de la campaña electoral	340	4,97
Mensajes con propuestas de acción política	778	11,37
Uso de eslóganes electorales	427	6,24
Total	6.845	100

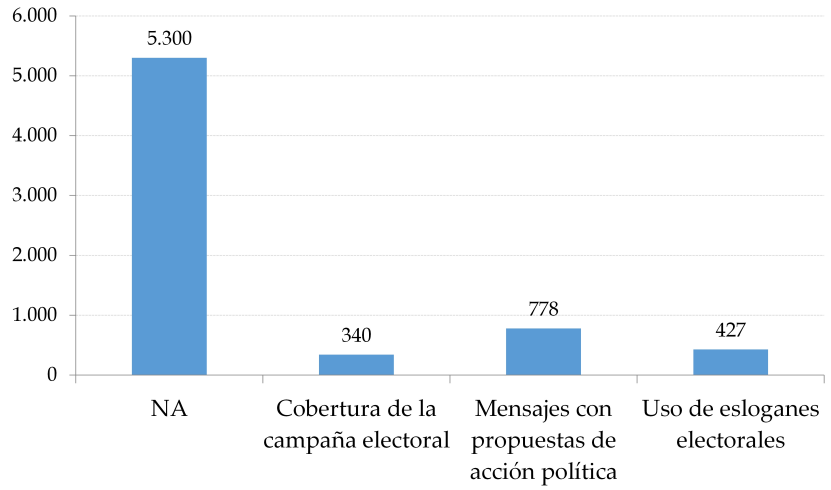


Figura 3.40: Histograma con la distribución de mensajes respecto al nivel de meta-comunicación.

nidos, ya sean propios, emitidos por un medio de comunicación o por un tercero distinto. También se puede ver gráficamente en la Figura 3.41.

Tabla 3.8: Distribución de mensajes según la referencialidad de contenidos.

Clase	Número de mensajes	%
Autorreferencial	265	3,87
El contenido referido publicado por un medio de comunicación	195	2,85
El contenido referido publicado por otro	299	4,37
Sin referencias a otros contenidos	6.086	88,91
Total	6.845	100

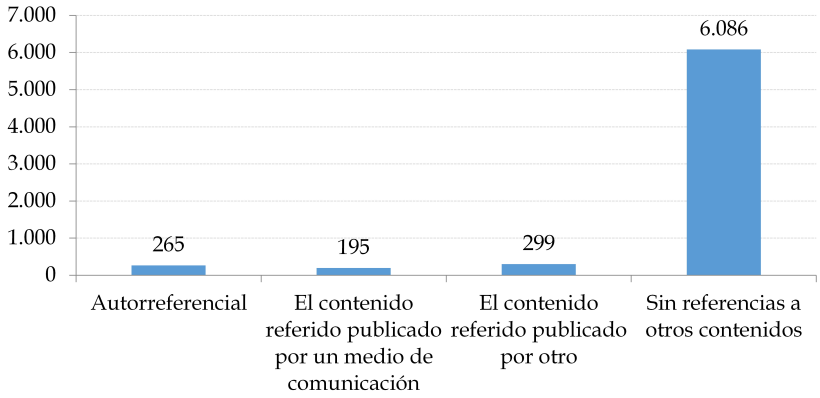


Figura 3.41: Histograma con la distribución de mensajes respecto a la referencialidad de contenidos.

Con respecto al nivel de personalización de los mensajes, las clases se distribuyen de tal modo que en su mayoría (dos de cada tres mensajes) no mencionan directamente a ningún candidato, ya sea propio o del adversario político. Aproximadamente uno de cada cinco mensajes habla del candidato propio, mientras que solo una décima parte hace alusión directa al candidato adversario.

Tabla 3.9: Distribución de mensajes según la personalización.

Clase	Número de mensajes	%
N/A	4.560	66,62
Menciona al candidato propio	1.456	21,27
Menciona al propio partido	672	9,82
Menciona a otros candidatos	157	2,29
Total	6.845	100

Como muestra la Tabla 3.10, el nivel de campaña negativa está en el entorno del 10 %, con 786 apariciones, y cuando se produce se ha referencia principalmente al partido político adversario en mayor medida frente al candidato adversario. La distribución de frecuencias se puede ver gráficamente en la Figura 3.43.

A continuación se describen las distribuciones de cada una de las clases asociadas a los marcos comunicacionales primario y secundario.

El principal marco primario utilizado en los mensajes es el responsabilidad en su modalidad de inmoralidad del oponente, seguido del marco de conflicto, que en total suponen una tercera parte de los marcos etiquetados, tal como

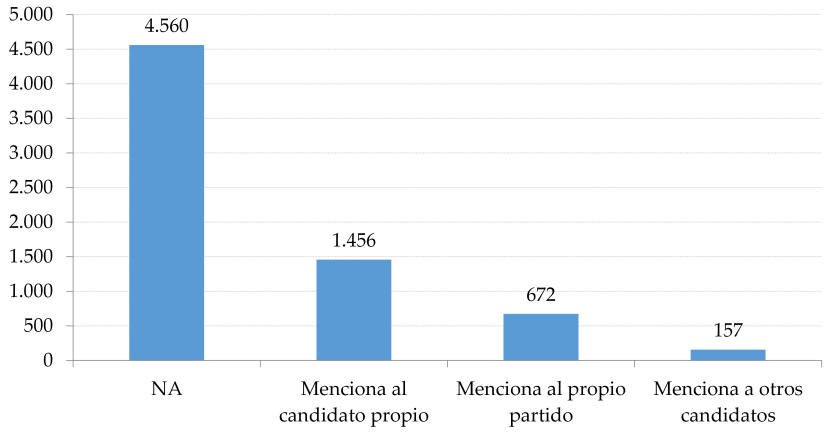


Figura 3.42: Histograma con la distribución de mensajes respecto a la personalización.

Tabla 3.10: Distribución de mensajes según el nivel de campaña negativa.

Clase	Número de mensajes	%
N/A	6.059	88,52
Mención a candidatos adversarios	228	3,33
Mención a partidos adversarios	558	8,15
Total	6.845	100

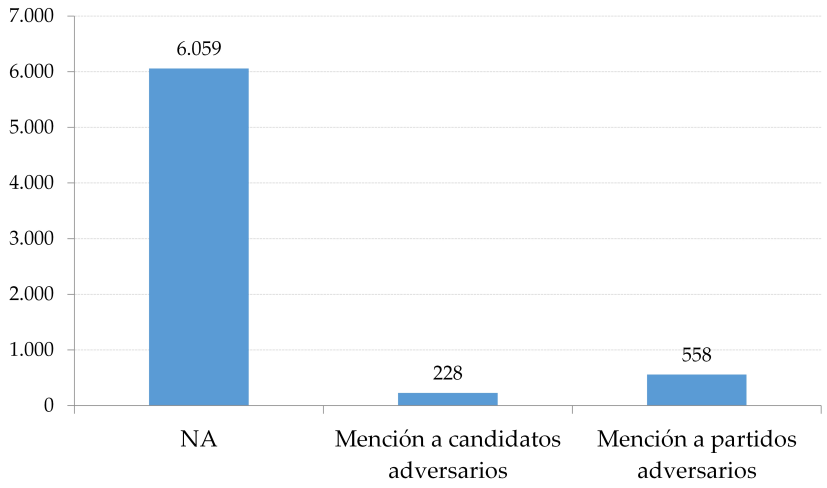


Figura 3.43: Histograma con la distribución de mensajes respecto a la clase de negatividad en la campaña.

recoge la Tabla 3.11. Estos datos se muestran gráficamente en la Figura 3.44.

Tabla 3.11: Distribución de mensajes según el marco primario, incluyendo el desglose por marco de responsabilidad.

Clase	Número de mensajes	%
N/A	3.000	43,83
Conflicto	1.001	14,62
Consecuencias económicas	49	0,72
Interés humano	373	5,45
Candidato como estratega	314	4,59
Candidato como héroe	370	5,41
Candidato como víctima	83	1,21
Candidato como populista	113	1,65
Inmoralidad del oponente	1.542	22,53
Total	6.845	100

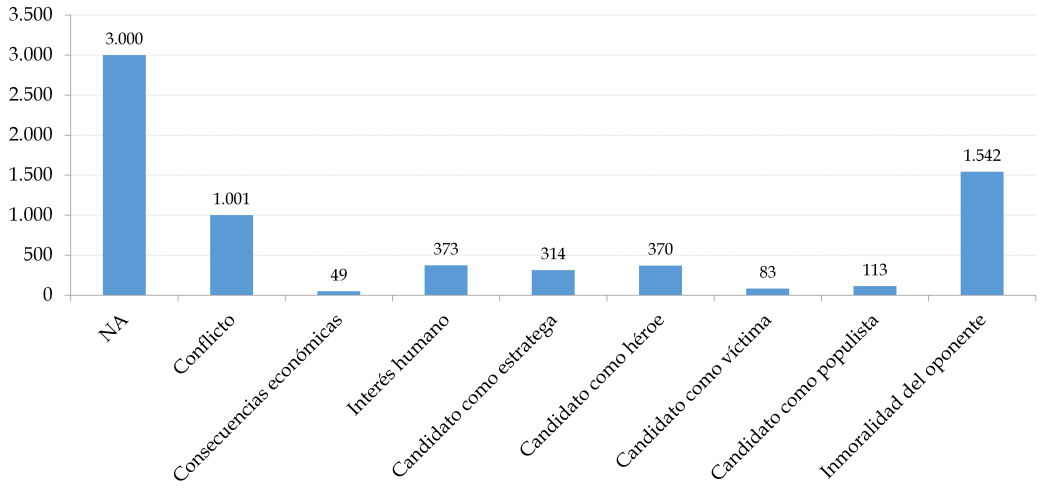


Figura 3.44: Histograma con la distribución de mensajes respecto al marco primario.

En la Tabla 3.11 se muestran los resultados del etiquetado del marco primario de los mensajes. El principal marco primario utilizado en los mensajes es el responsabilidad, y en particular en su modalidad de inmoralidad del oponente, seguido del marco de conflicto que en total suponen una tercera parte de los marcos etiquetados. Estos datos se muestran gráficamente en la Figura 3.45.

Una vez repasados los resultados del proceso de etiquetado de los mensajes,

Tabla 3.12: Distribución de mensajes según el marco secundario, incluyendo el desglose por marco de responsabilidad.

Clase	Número de mensajes	%
N/A	6.327	92,43
Conflicto	101	1,48
Consecuencias económicas	27	0,39
Interés humano	101	1,48
Candidato como estratega	18	0,269
Candidato como héroe	24	0,35
Candidato como víctima	0	0,00
Candidato como populista	32	0,47
Inmoralidad del oponente	215	3,14
Total	6.845	100

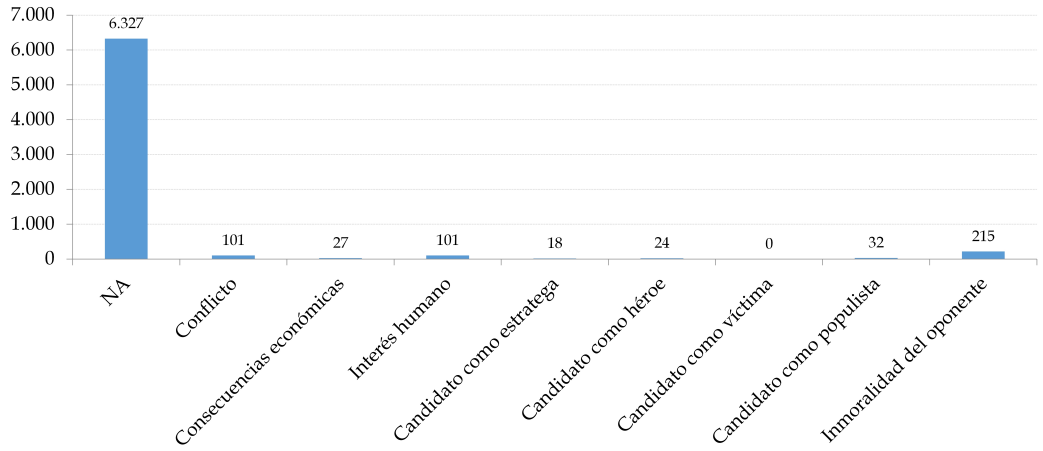


Figura 3.45: Histograma con la distribución de mensajes respecto al marco secundario.

pasamos a detallar los resultados del proceso de etiquetado de las conversaciones.

3.4.3.2 **Conversaciones políticas**

Recordemos que las conversaciones se han etiquetado según los siguientes tipos de clasificación, a tenor de los correspondientes criterios: la función informativa, la función movilizadora, el nivel de personalización, el nivel de meta.comunicación, y los marcos primario y secundario, además del nivel de cam-

paña negativa.

Atendiendo a la función informativa, las conversaciones se distribuyen principalmente sin enlaces, seguidos de aquellas que tienen enlaces a terceros y por último las que tienen enlaces propios, según se muestra en la Figura 3.46.

Tabla 3.13: Distribución de conversaciones con respecto a la función informativa.

Clase	Número de conversaciones	(%)
Función informativa sin enlaces	34	43,04
Función informativa con enlaces propios	18	22,78
Función informativa con enlaces a terceros	27	34,18
Total	79	100

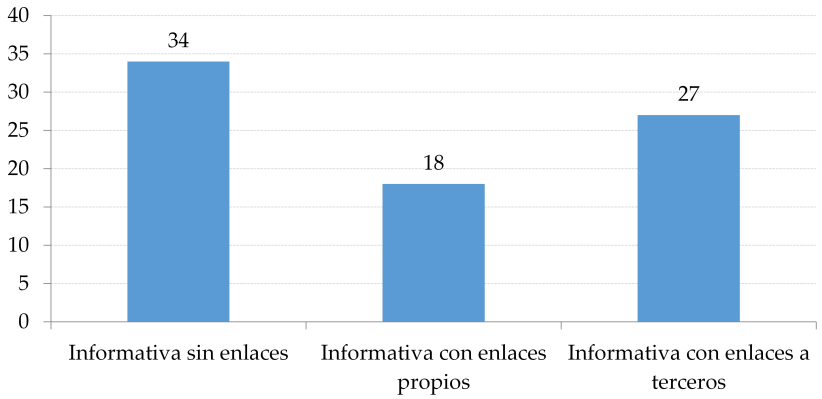


Figura 3.46: Histograma con la distribución de conversaciones políticas respecto a la función informativa.

Por su parte, siguiendo la clasificación de la función movilizadora, las conversaciones siguen la distribución que se muestra en la Tabla **??**. Según estos datos, las conversaciones no desempeñan una función de movilización (menos de una tercera parte son movilizadores), y solo una cuarta parte de ellos hacen referencia al voto.

En lo que se refiere al análisis de la meta-comunicación, la distribución de las conversaciones en torno a la clasificación del nivel de meta-comunicación señalan que tan solo una quinta parte de los mensajes presenta un nivel de meta-comunicación de campaña significativo en términos de la cobertura del proceso electoral, mensajes con propuestas de acción política o bien el uso de

Tabla 3.14: Distribución de conversaciones respecto a la función movilizadora.

Clase	Número de mensajes	%
N/A	53	67,09
Movilización hacia el voto	6	7,59
Movilización hacia otros tipos de acción	20	25,32
Total	79	100

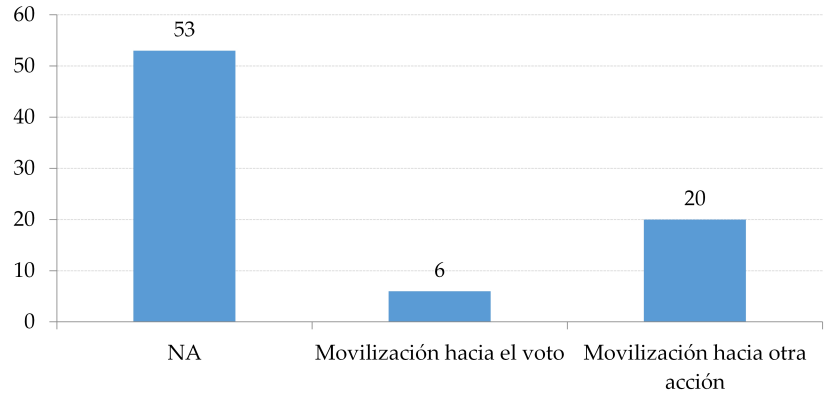


Figura 3.47: Histograma con la distribución de conversaciones respecto a la función movilizadora.

eslóganes electorales. La Tabla 3.15 muestra la distribución de los resultados. En la figura 3.48 se representa gráficamente esta distribución.

Tabla 3.15: Distribución de conversaciones respecto al nivel de meta-comunicación.

Clase	Número de mensajes	%
N/A	29	36,71
Cobertura de la campaña electoral	22	27,85
Mensajes con propuestas de acción política	23	29,11
Uso de eslóganes electorales	5	6,33
Total	79	100

Las conversaciones muestran una componente de campaña negativa muy reducida si se observa su distribución de frecuencias (Figura 3.49), donde apenas un 10 % de las conversaciones se plantean en términos de negatividad frente a los adversarios políticos.

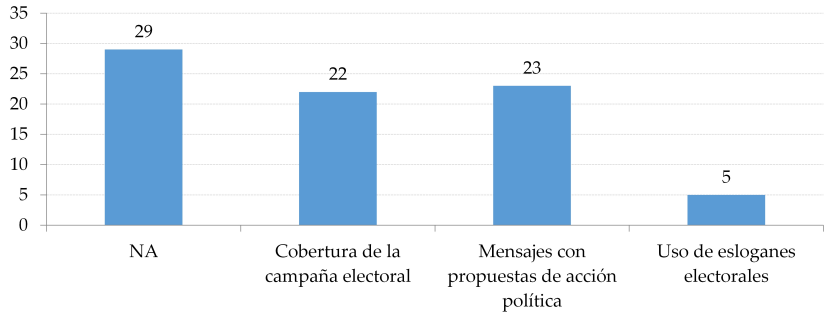


Figura 3.48: Histograma con la distribución de conversaciones respecto al nivel de meta-comunicación.

Tabla 3.16: Distribución de conversaciones respecto al nivel de personalización.

Clase	Número de mensajes	%
N/A	70	88,61
Mención a candidatos adversarios	8	10,13
Mención a partidos adversarios	1	1,27
Total	79	100

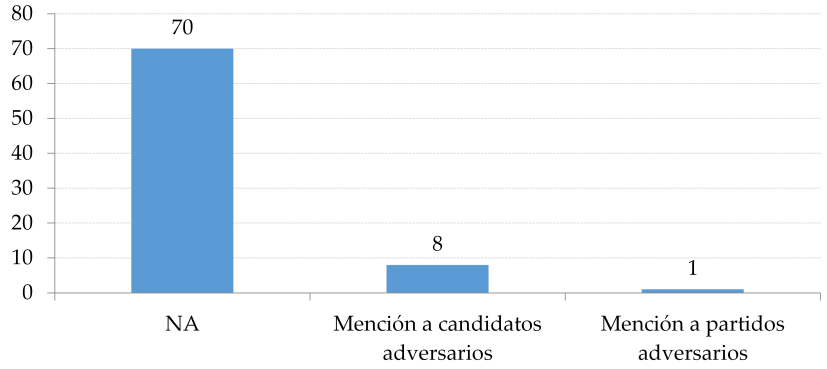


Figura 3.49: Histograma con la distribución de conversaciones respecto al nivel de campaña negativa.

A continuación se describen las distribuciones de cada una de las clases asociadas a los marcos comunicacionales primario y secundario (Véase la Tabla 3.17). En primera instancia, los marcos primarios o predominantes que se activan con más frecuencia son los de responsabilidad (que en total suman el 39,24 % de los casos), seguidos del conflicto y del interés humano.

Tabla 3.17: Distribución de conversaciones respecto al marco primario, incluyendo el desglose por marco de responsabilidad.

Clase	Número de mensajes	%
N/A	17	21,52
Conflicto	15	18,99
Consecuencias económicas	5	6,33
Interés humano	11	13,92
Candidato como estratega	8	10,13
Candidato como héroe	2	2,53
Candidato como víctima	0	0,00
Candidato como populista	9	11,39
Inmoralidad del oponente	12	15,19
Total	6.845	100

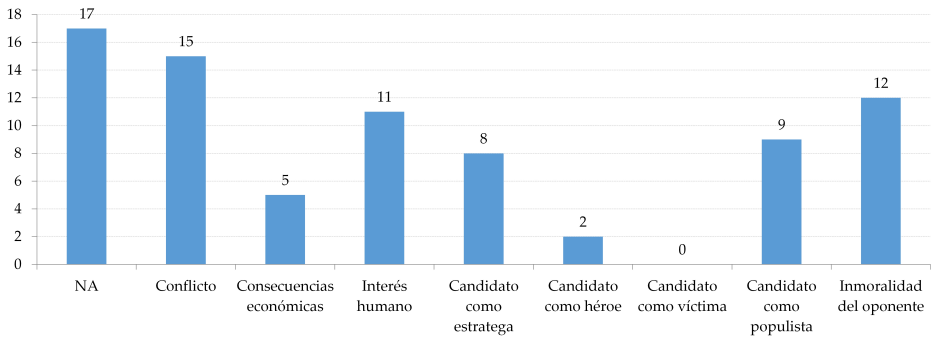


Figura 3.50: Histograma con la distribución de conversaciones respecto a la clase de marco primario.

Por su parte, la distribución de frecuencias de las conversaciones según el marco secundario queda recogida en la Tabla 3.18. La frecuencia de los marcos comunicacionales es menor que en el marco primario, con una presencia relevante del marco de responsabilidad que, en su conjunto, supone aproximadamente un 25 % de las apariciones totales.

Tabla 3.18: Distribución de conversaciones respecto al marco secundario, incluyendo el desglose por marco de responsabilidad.

Clase	Número de mensajes	%
N/A	48	60,76
Conflicto	2	2,53
Consecuencias económicas	1	1,27
Interés humano	8	10,13
Candidato como estratega	0	0.00
Candidato como héroe	1	1,27
Candidato como víctima	0	0,00
Candidato como populista	8	10,13
Inmoralidad del oponente	11	13,92
Total	6.845	100

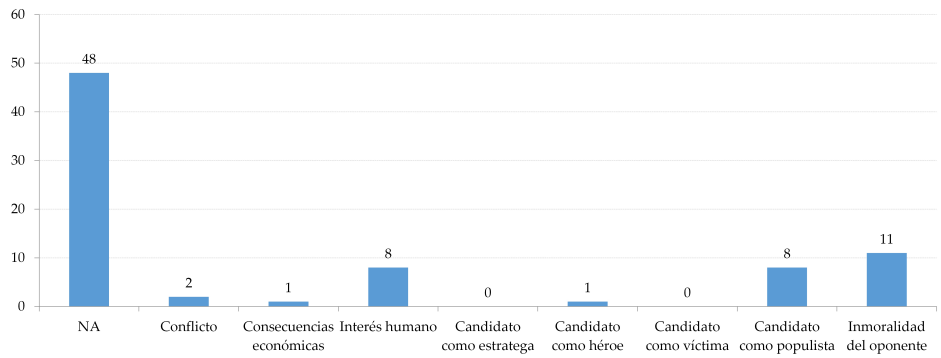


Figura 3.51: Histograma con la distribución de conversaciones respecto a la clase de marco secundario.

3.5 Adaptación de los datos

En esta sección, se describen las características que se han incorporado a los análisis, tanto de los mensajes individuales como de las conversaciones políticas en los muros digitales de los partidos.

3.5.1 Características extraídas de los mensajes

En el caso de los comentarios, hemos dividido estas características en 2 categorías diferentes: propiedades estadísticas y propiedades lingüísticas.

3.5.1.1 **Propiedades estadísticas**

La categoría de propiedades estadísticas dispone de varias características, las cuales hemos empleado para la construcción del conjunto final de datos.

- **Cuerpo del mensaje.** Hemos utilizado la información contenida en el cuerpo del mensaje. A fin de poder procesar la información de que contienen los mensajes de manera automática, se ha hecho uso del modelo de recuperación de información o IR, del inglés *Information Retrieval*). Un modelo IR puede ser definido como una 4-tupla de la manera $[\mathscr{C}, \mathscr{Q}, \mathscr{F}, \mathscr{R}(q_i, c_j)]$ [BYRN99b] donde:
  - $\mathscr{C}$ es el conjunto de representaciones de mensajes.
  - $\mathscr{Q}$ es el conjunto de consultas del usuario.
  - $\mathscr{F}$ es el marco de trabajo para modelar mensajes, consultas y sus relaciones.
  - $\mathscr{R}(q_i, c_j)$ es una función de clasificación que asocia un número real con una consulta q_i ($q_i \in \mathscr{Q}$) y la representación de un mensaje c_j, por lo que ($c_j \in \mathscr{C}$).

  Dado $\mathscr{C}$ como conjunto de comentarios c, $\{c : \{t_1, t_2, ...t_n\}\}$, donde cada uno de los comentarios comprende n términos $t_1, t_2, \ldots, t_n$, se ha definido el peso $w_{i,j}$ como el número de veces que aparece el término t_i en el comentario c_j. Si $w_{i,j}$ no está presente en c, el peso $w_{i,j} = 0$. Por lo tanto, un mensaje c_j puede ser representado como un vector de pesos $\vec{c_j} = (w_{1,j}, w_{2,j}, ...w_{n,j})$.

  Sobre la base de esta formalización, los sistemas IR normalmente emplean el modelo espacio vectorial (VSM, de los términos ingleses *Vector Space Model*) [BYRN99b], que algebraicamente representa documentos (en nuestro caso mensajes políticos) como vectores en un espacio multidimensional.

  Este espacio se compone, únicamente, de las intersecciones de los ejes positivos. Los comentarios son representados por una matriz *término-por-comentario*, donde el elemento (i, j) ilustra la asociación entre el término i y el comentario j. Esta asociación refleja la aparición del término i en el comentario j. Los términos pueden representar diferentes unidades textuales (por ejemplo, palabras o frases) y también pueden ser ponderados individualmente, a través de un sistema de pesos, permitiendo a los términos convertirse en más o menos importantes dentro de un comentario dado o dentro de una colección de comentarios $\mathscr{C}$ al completo.

Como esquema de pesos se ha utilizado el esquema de ponderación *frecuencia del término-frecuencia inversa del documento*, de la voz anglosajona *Term Frequency-Inverse Document Frequency* o TF-IDF [SM83], según la cual el peso del término i en el mensaje j, $weight(i,j)$, se define como:

$$weight(i,j) = tf_{i,j} \cdot idf_i \tag{3.1}$$

siendo la frecuencia del término $tf_{i,j}$:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3.2}$$

donde $n_{i,j}$ es el número de veces que el término $t_{i,j}$ aparece en el mensaje c y $\sum_k n_{k,j}$ es el número total de términos en el propio mensaje c.

Por otra parte, la frecuencia inversa del documento, se define como:

$$idf_i = \frac{|\mathscr{C}|}{|\mathscr{C} : t_i \in c|} \tag{3.3}$$

donde $|\mathscr{C}|$ es el número total de mensajes y $|\mathscr{C} : t_i \in c|$ es el número de mensajes que contienen el término t_i.

Tras definir el modelo de representación de la información y el esquema de ponderación, resulta necesario establecer los términos textuales que son objeto de ponderación y recuperación.

Tanto en el caso de los mensajes políticos como en el de los post iniciales de cada conversación se han utilizado los mismos esquemas. En primer lugar el *término textual* utilizado es el token o simplemente término y, como segunda opción, se ha utilizado el enfoque *n-grama*. Dado el tamaño variable de un n-grama como término, esto es, subsecuencia de n elementos consecutivos en una secuencia dada, se han establecido distintas dimensiones, partiendo de un tamaño mínimo de $n = 1$, y un máximo establecido en $n = 3$ para el contenido de cada mensaje.

Con el fin de construir el modelo, también hemos suprimido las palabras de parada (del inglés *stop-words*)[WS92], las cuales son palabras desprovistas de contenido o palabras las cuales su aparición en el idioma empleado es muy habitual, con lo cual no aportan información alguna (por ejemplo, un, el, es). Estas palabras no proporcionan ninguna información

semántica y añaden ruido al modelo [SM83]. Hemos utilizado una lista de palabras de parada externa de términos en castellano[1].

3.5.1.2 **Propiedades lingüísticas**

Se han incorporado al análisis determinadas características lingüísticas que se han extraído a partir del texto del mensaje. Para ello se ha hecho uso de la tecnología FreeLing en su versión 3.1 con el fin de realizar un etiquetado de tipología de palabra en la oración (*PoS* o *Part of Speech* en inglés)[2].

Con este etiquetado sintáctico se obtiene información relevante del tipo de palabras usadas en los mensajes. De este modo, el uso de una gran cantidad de verbos puede estar expresando una determinada orientación hacia la acción, mientras que si aparecen muchos adjetivos se puede estar tratando de un texto más descriptivo u opinativo.

- Número de verbos en el texto del mensaje.
- Número de nombres en el texto del mensaje.
- Número de adjetivos en el texto del mensaje.
- Número de números en el texto del mensaje.
- Número de adverbios en el texto del mensaje.
- Número de nombres en el texto del mensaje.
- Número de conjunciones en el texto del mensaje.
- Número de pronombres en el texto del mensaje.
- Número de signos de puntuación en el texto del mensaje.
- Número de interjecciones en el texto del mensaje.
- Número de determinantes en el texto del mensaje.
- Número de adposiciones en el texto del mensaje.
- Número de preposiciones en el texto del mensaje.

[1] La lista de palabras de parada puede descargarse en: `http://paginaspersonales.deusto.es/isantos/resources/stopwords.txt`

[2] Disponible en `http://nlp.lsi.upc.edu/freeling/`

3.5.2 Características extraídas de las conversaciones

A partir de las propiedades del conjunto de datos de las conversaciones se han generado determinadas características que se incluyen para su análisis final. A fin de estructurar su naturaleza u origen, se detallan a continuación organizadas en distintos tipos: (i) propiedades generales, (ii) propiedades estadísticas, (iii) propiedades comunicacionales de la conversación, (iv) propiedades del *post* inicial y (v) propiedades del flujo de mensajes.

3.5.2.1 Propiedades generales

Estas propiedades son propias y definitorias de cada una de las conversaciones políticas.

- **Identificador conversación**. Es un identificador numérico que señala biunívocamente cada una de las conversaciones que tienen lugar.
- **Perfil emisor**. Identifica el partido que inicia la conversación en su muro; en nuestro caso podrán tomar el valor PP, PSOE y Podemos.
- **Fecha inicio**. Momento en el que se origina la conversación.
- **Grado de aprobación**. Es el número de veces que los usuarios han pulsado la funcionalidad "*Me gusta*".
- **Grado de compartición**. Es el número de veces que los usuarios han compartido el mensaje inicial de la conversación.

3.5.2.2 Propiedades estadísticas

A partir de cada conversación se desprenden distintos atributos que sintetizan elementos coyunturales de cada conversación analizada.

- **Nivel de respuesta**. Número de mensajes de respuesta que se han producido en la conversación a partir del *post* inicial.
- **Partido emisor**. Identifica el partido que inicia la conversación en su muro; en nuestro caso podrán tomar los valores PP, PSOE, Podemos.
- **Número de conversadores**. Indica el número total de perfiles distintos que han participado en la conversación.

- **Alcance total**. Mide el grado en que un mensaje inicial ha producido efectos, bien sea generando un mensaje nuevo, una compartición o una muestra de acuerdo.

3.5.2.3 **Propiedades comunicacionales de la conversación**

Tras etiquetar y categorizar cada una de las conversaciones, cada una de ellas cuenta con atributos asignados. Las propiedades que se enuncian seguidamente son las correspondientes a este etiquetado, y que se han descrito previamente en la sección 3.4.

- Función informativa.
- Función movilizadora.
- Nivel de personalización.
- Nivel de meta-comunicación.
- Grado de debate político.
- Referencialidad de los contenidos.
- Marco primario.
- Marco secundario.
- Negatividad de campaña.

3.5.2.4 **Propiedades del post inicial**

El mensaje inicial o post aporta distintas características al conjunto de características de la conversación, a saber:

- **Cuerpo del mensaje o post inicial**. Representa el contenido del texto del mensaje inicial.
- **Total de enlaces**. Contabiliza el número de hiperenlaces externos que aparecen en el cuerpo del mensaje inicial.
- **Total imágenes**. Representa el número de imágenes existentes dentro del cuerpo del *post* inicial.

- **Total vídeos**. De manera análoga a las dos características anteriores, pero para vídeos accesibles a través del sitio social *YouTube*.
- **Total etiquetas**. Recoge el número total de etiquetas en forma de *hastag* que contiene el cuerpo del mensaje.
- **Grado de difusión del *post* inicial**. Número de veces que el post inicial ha sido compartido por algún usuario.
- **Grado de aceptación del *post* inicial**. Número de veces que los usuarios han señalado que les gusta a través de la función "*Me gusta*".

3.5.2.5 **Propiedades del flujo de mensajes de respuesta**

Además de lo anterior, se han utilizado características obtenidas a partir del conjunto de mensajes que se han generado tras el mensaje inicial. Incorpora el grado de debate político de los mensajes, descrito previamente en 3.4.

- Número de mensajes con función informativa con enlaces a terceros.
- Número de mensajes con función informativa con enlaces propios.
- Número de mensajes con función informativa sin enlaces.
- Número de mensajes con función movilizadora hacia el voto.
- Número de mensajes con función movilizadora hacia otro tipo de movilización.
- Número de mensajes autorreferenciales.
- Número de mensajes que referencian a contenido publicado por un medio de comunicación.
- Número de mensajes que referencian a contenido publicado por otro.
- Número de mensajes sin referencias a otros contenidos.
- Número de mensajes en los que se menciona al candidato propio.
- Número de mensajes en los que se menciona al partido propio.
- Número de mensajes en los que se menciona a otros candidatos.

- Número de mensajes negativos de campaña que se refieren al candidato adversario.
- Número de mensajes negativos de campaña que se refieren al partido adversario.
- Número de mensajes con nivel de meta-comunicación de cobertura de campaña.
- Número de mensajes con nivel de meta-comunicación de uso de eslóganes.
- Número de mensajes con nivel de meta-comunicación de uso propuestas de acción política.
- Número de mensajes irrelevantes.
- Número de mensajes reafirmativos.
- Número de mensajes controvertidos.
- Número de mensajes polémico.
- Número de mensajes ofensivos o insultos.
- Número de mensajes cuyo marco primario es de conflicto.
- Número de mensajes cuyo marco primario es de consecuencias económicas.
- Número de mensajes cuyo marco primario es de interés humano.
- Número de mensajes cuyo marco primario es de responsabilidad del tipo candidato como estratega.
- Número de mensajes cuyo marco primario es de responsabilidad del tipo candidato como héroe.
- Número de mensajes cuyo marco primario es de responsabilidad del tipo candidato como víctima.
- Número de mensajes cuyo marco primario es de responsabilidad del tipo candidato como populista.
- Número de mensajes cuyo marco primario es de responsabilidad del tipo inmoralidad del oponente.

- Número de mensajes cuyo marco secundario es de conflicto.
- Número de mensajes cuyo marco secundario es de consecuencias económicas.
- Número de mensajes cuyo marco secundario es de interés humano.
- Número de mensajes cuyo marco secundario es de responsabilidad del tipo candidato como estratega.
- Número de mensajes cuyo marco secundario es de responsabilidad del tipo candidato como héroe.
- Número de mensajes cuyo marco secundario es de responsabilidad del tipo candidato como víctima.
- Número de mensajes cuyo marco secundario es de responsabilidad del tipo candidato como populista.
- Número de mensajes cuyo marco secundario es de responsabilidad del tipo inmoralidad del oponente.

3.5.3 **Discusión de los resultados**

Una vez adaptados los datos de las 79 conversaciones políticas y los 6.845 mensajes, se han creado 34 ficheros ARFF [1] [HDW94], que contienen la relación de los atributos de cada uno de estos tipos de entidades.

Para los mensajes, se genera un archivo que contiene la representación vectorial de las características del mensaje para cada una de las clasificaciones a las que se somete (9 clasificadores), y uno por cada tipo de modelo de representación vectorial (*VSM*) que se utilice para representar el cuerpo del mensaje (n-gramas y palabras). En total, se generan 18 ficheros para realizar las distintas clasificaciones de los mensajes.

Análogamente, para las conversaciones, se elabora un fichero por cada una de las representaciones vectoriales de sus características, uno por cada uno de los tipos de representación del espacio vectorial (*VSM*) que representa el cuerpo del mensaje inicial (n-gramas y palabras), y uno por cada uno de las 7 clasificaciones de las conversaciones. En total se generan 14 ficheros ARFF para clasificaciones.

[1] ARFF: Attribute Relation File Format

El empleo de los ficheros ARFF viene motivado por el uso de una herramienta de aprendizaje automático denominada WEKA (acrónimo del inglés *Waikato Environment for Knowledge*) [G+95], que limita el formato de los ficheros de entrada, sin menoscabo de otras consideraciones o efectos secundarios.

Recordemos que el conjunto de mensajes se ha etiquetado a tenor de las 10 clasificaciones siguientes:

- Función informativa.
- Función movilizadora.
- Nivel de personalización.
- Nivel de meta-comunicación.
- Grado de debate político.
- Grado de contribución
- Referencialidad de contenidos.
- Nivel de campaña negativa.
- Marco primario.
- Marco secundario.

Por su parte, en el conjunto de datos de conversaciones, el etiquetado se ha realizado en base a 7 categorías diferentes:

- Función informativa.
- Función movilizadora.
- Nivel de personalización.
- Nivel de meta-comunicación.
- Nivel de campaña negativa.
- Marco primario.
- Marco secundario.

Con respecto al *Vector Space Model* o (VSM), se han incorporado 2 posibles enfoques:

- La palabra o *token*.
- El *n-grama*.

En total, la aplicación del VSM para cada una de las categorías – 10 para los mensajes y 7 para las conversaciones – de cada uno de los conjuntos de datos – mensajes políticos y conversaciones políticas – ha implicado la generación de un total de 34 archivos ARFF; 20 de ellos asociados al análisis de los comentarios, y que se corresponden con la siguiente enumeración:

- La función informativa del mensaje con el VSM basado en palabras.
- La función informativa del mensaje con el VSM basado en *n-gramas*.
- La función movilizadora del mensaje con el VSM basado en palabras.
- La función movilizadora del mensaje con el VSM basado en *n-gramas*.
- El nivel de personalización del mensaje con el VSM basado en palabras.
- El nivel de personalización del mensaje con el VSM basado en *n-gramas*.
- El nivel de meta-comunicación del mensaje con el VSM basado en palabras.
- El nivel de meta-comunicación del mensaje con el VSM basado en *n-gramas*.
- El grado de contribución del mensaje con el VSM basado en palabras.
- El grado de contribución del mensaje con el VSM basado en *n-gramas*.
- El grado de debate político del mensaje con el VSM basado en palabras.
- El grado de debate político del mensaje con el VSM basado en *n-gramas*.
- La referencialidad de contenidos del mensaje con el VSM basado en palabras.
- La referencialidad de contenidos del mensaje con el VSM basado en *n-gramas*.

- El nivel de campaña negativa del mensaje con el VSM basado en palabras.
- El nivel de campaña negativa del mensaje con el VSM basado en *n-gramas*.
- El marco primario del mensaje con el VSM basado en palabras.
- El marco primario del mensaje con el VSM basado en *n-gramas*.
- El marco secundario del mensaje con el VSM basado en palabras.
- El marco secundario del mensaje con el VSM basado en *n-gramas*.

Adicionalmente, para las conversaciones políticas se han producido otros 14 archivos ARFF con un propósito análogo:

- La función informativa de la conversación con el VSM basado en palabras.
- La función informativa de la conversación con el VSM basado en *n-gramas*.
- La función movilizadora de la conversación con el VSM basado en palabras.
- La función movilizadora de la conversación con el VSM basado en *n-gramas*.
- El nivel de personalización de la conversación con el VSM basado en palabras.
- El nivel de personalización de la conversación con el VSM basado en *n-gramas*.
- El nivel de meta-comunicación de la conversación con el VSM basado en palabras.
- El nivel de meta-comunicación de la conversación con el VSM basado en *n-gramas*.
- El nivel de campaña negativa de la conversación con el VSM basado en palabras.
- El nivel de campaña negativa de la conversación con el VSM basado en *n-gramas*.
- El marco primario de la conversación con el VSM basado en palabras.

- El marco primario de la conversación con el VSM basado en *n-gramas*.
- El marco secundario de la conversación con el VSM basado en palabras.
- El marco secundario del mensaje con el VSM basado en *n-gramas*.

Una vez realizado el proceso de aplicación de VSM a ciertas características de nuestros conjuntos de datos y la creación de los archivos, la Tabla 3.19 muestra los resultados logrados para los diferentes conjuntos de datos.

Tabla 3.19: Número de características para cada conjunto de datos.

Conjunto de datos	**Número de características**	
	Con n-gramas	**Con palabras**
Mensajes	*1.024*	*1.043*
Conversaciones	*7.806*	*1.175*

Se puede observar el gran aumento de características debido a la aplicación tanto del enfoque de palabra basado en VSM como el enfoque de n-grama basado también en VSM, sobre el cuerpo del mensaje y sobre el cuerpo del *post* o mensaje inicial de la conversación. La principal razón de este incremento es el desglose de ambos textos en relación con las palabras más relevantes y en los *unigramas*, *bigramas* y *trigramas*, esto es, los N-gramas con valores de $N = 1, 2, 3$.

La utilización de la técnica basada en VSM y el esquema de ponderación TF-IDF está basada en que en nuestro caso nos ofrece una serie de ventajas que favorecen el modelado.

Para la técnica TF-IDF, el procedimiento de TF clasifica los términos por mayor importancia, debido a su mayor frecuencia en el cuerpo de un mensaje, lo cual nos permite filtrar los términos con menor relevancia dentro del propio texto del mensaje. No obstante, ante la hipotética restricción de mensajes con tamaño elevado, se constata que los textos que aparecen en los mensajes no verifican esta condición, por lo que no resulta necesario normalizar dichos textos, que es la técnica a emplear en tal caso.

Así mismo, para la técnica IDF resulta adecuada su apliación en tanto en cuanto a través de ella se obtienen los términos que más relevantes son en el texto a partir del nivel de singularidad que aportan, esto es, discriminan más en la medida en que son menos frecuentes. Utilizando los términos que discriminan con mayor intensidad se define el concepto de *escasez* de un término.

Dicho lo anterior, debido a la utilización conjunta de ambos procedimientos, TF-IDF, es posible conseguir una recuperación ordenada, de tal modo que si el

texto de un mensaje es devuelto antes que otro, lo es porque que dicho texto es más relevante que el otro.

Aunque por otra parte, se puede dar el caso del acoplamiento parcial, el cual indica que el comentario más relevante no tiene por qué contener todos los términos de la consulta. Pero no tiene por qué haber ningún comentario de este tipo, aunque hubiera comentarios relevantes. Por último, el esquema TF-IDF es bastante intuitivo y a su vez, asume la independencia de términos (más conocido como *bag of words*).

Por otra parte, gracias al empleo de *n-gramas*, en nuestro caso *unigramas*, *bigramas* y *trigramas*, podríamos incrementar el contexto del estudio aumentando el tamaño de *n*, a fin de experimentar con nuevos indicadores a la espera de mejorar los resultados. Al *tokenizar* los comentarios en *n-gramas*, en vez de palabras, se consigue que el modelo sea independiente del dominio del idioma, ya que no integra ningún tipo de recurso concreto sobre ellos. Al contrario, el empleo de instancias a nivel de *n-grama* constituye en sí mismo un mecanismo de normalización de términos que permite, además, trabajar con una gran variedad de lenguas sin procesamiento alguno [MM04a, MM04b, RW98].

3.6 Sumario

A lo largo de este capítulo se han abordado el detalle del procesamiento de los datos. Así, en primer lugar se ha presentado la arquitectura general del sistema propuesto para dar respuesta al diseño de la investigación. Seguidamente se ha realizado una descripción del sitio web de interacción social *Facebook* que sirve de fuente primaria de los datos para nuestra investigación. Con el fin de acercarnos más al objeto de la investigación, se han explicitado los principales rasgos que explican la idoneidad de esta red social.

En la etapa de etiquetado se ha procedido, en primer lugar a describir el procesamiento conducente a la recolección de los datos de manera automática. Seguidamente se han detallado los modelos de datos que permiten representar cada uno de los elementos a observar (mensajes y conversaciones políticas), para posteriormente presentar los distintos sistemas de clasificación que se han empleado para los mensajes publicado en la red social, así como al conjunto de las discusiones políticas que son objeto de estudio.

Para finalizar con las tareas, la adaptación de datos ha permitido extraer características en cada uno de los conjuntos de datos. De este modo, la representación de los mensajes se ha enriquecido a través de dos tipos de características – estadísticas y lingüísticas – que ha sido posible generar a partir de la

información de base que se ha capturado de forma automática.

Del mismo modo, la información recogida para las conversaciones ha posibilitado que podamos disponer de nuevas características de distinto tipo: (i) información general de la conversación, (ii) propiedades estadísticas, (iii) propiedades comunicacionales de la propia conversación, (iv) propiedades relacionadas con el *post* inicial y, finalmente (v) las originadas a partir del flujo de mensajes de respuesta.

El capítulo ha concluido con la exposición de los resultados de la etapa de etiquetado y su análisis. Esta etapa es previa y necesaria para la posterior tarea de clasificación automática, y que se presenta en el capítulo siguiente.

«Sobre lo que no podemos hablar
hemos de guardar silencio»

Ludwig Wittgenstein (1889-1951)

CAPÍTULO

4

Categorización automática de mensajes y conversaciones políticas

La emergencia de las nuevas redes de interacción social *online* como un nuevo espacio en el que las personas realizan intercambio de opiniones, aportaciones de información y compartición de recursos cada vez en más formatos y con mayor calidad ha suscitado el interés de empresas, organizaciones e instituciones por conocer cómo es y en qué consiste esta nueva dinámica, a fin de preservar su imagen o reputación, o bien de obtener resultados económicos, sociales, políticos o institucionales a partir de ellas.

La doble disponibilidad de tecnologías que permiten, por un lado el acceso automático a dicha información y por el otro lado, para el procesamiento de dichos datos en un tiempo razonable y con resultados significativos ha venido suponiendo un notable avance en el terreno de la clasificación automática de información representada de manera digital.

Debido a este creciente aumento, existe una disponibilidad de información digital, lo que ha permitido que la categorización automática sea algo accesible y por lo tanto, el desarrollo de diferentes técnicas para su implementación. Como se ha visto en capítulos anteriores, la clasificación automática de texto consiste en un conjunto de algoritmos, técnicas y sistemas capaces de asignar

un documento a una o varias categorías o grupos de documentos, construidos según algún tipo de vinculación o afinidad. Para ello se emplean técnicas de Aprendizaje Automático (ML, *Machine Learning*) y de Procesamiento de Lenguaje Natural (NLP, *Natural Language Processing*).

En la presente investigación, es la red social *Facebook* el entorno digital sobre el que se realiza esa observación, y en particular, el terreno de los partidos políticos. En este sitio de interacción social los partidos publican mensajes a los que cualquier persona conectada a dicho partido puede responder con texto, señalar que está de acuerdo o muestra cierto agrado, o bien compartirlo con sus allegados en la red. Se establece así una interacción multilateral con distintos efectos y consecuencias sobre el conjunto de la red y que puede tener su traslación al mundo *real*, esto es, más allá de lo meramente digital.

Con el objeto de abordar este tema, a lo largo de este capítulo estableceremos:

- Un nuevo método basado en aprendizaje automático para categorizar los mensajes publicados y las conversaciones políticas que tienen lugar en este sitio de interacción social.

- Una vVerificación de que este método es capaz de clasificar con eficacia los aspectos clave de los mensajes que los ciudadanos emiten, así como de las conversaciones que se construyen a través de dichos mensajes.

El resto del capítulo se organiza como sigue. En la sección 4.1 se revisan los distintos métodos de aprendizaje supervisado utilizados en esta investigación y los principales clasificadores que se utilizan. A continuación, la sección 4.2 recoge el proceso en el que, a partir de las técnicas citadas anteriormente, se valida experimentalmente sendos modelos de clasificación de mensajes y conversaciones políticas. Tras esta etapa en la que se presentan los resultados, la sección 4.3 los analiza con el fin de obtener conclusiones claras sobre la clasificación de mensajes y conversaciones en el sitio de interacción social *Facebook*. Para finalizar, la sección 4.4 sintetiza las contribuciones alcanzadas a lo largo de este capítulo.

4.1 El aprendizaje automático

Una de las definiciones formales sobre el significado de los términos aprendizaje automático es la acuñada por Mitchell [Mit97], según la cual:

> «*Un programa de ordenador se dice que aprende de la experiencia E con respecto a una clase de tareas T y una medida de rendimiento P, si su rendimiento en tareas de T, según la medida P, mejora con la experiencia E.*»

Veamos un ejemplo: aprender a reconocer la escritura manual por parte de una máquina. En este caso:

- **Tarea (T).** Esta tarea supone reconocer y clasificar palabras manuscritas en una imagen.
- **Medida de rendimiento (P).** Hace referencia al porcentaje de palabras reconocidas correctamente.
- **Experiencia (E).** Se refiere a la base de datos de imágenes de palabras manuscritas, ya clasificadas.

El aprendizaje automático (ML, del inglés *Machine Learning*) es el área que se ocupa de estudiar la problemática sobre el aprendizaje por parte de las máquinas. Set trata de es un campo derivado de la inteligencia artificial o IA (AI, del inglés *Artificial Intelligence*).

La eficacia de las técnicas de aprendizaje automático ha sido probada en ámbitos diversos, y con resultados generalmente satisfactorios, lo que ha incentivado nuevos avances, tanto en cuanto a la sofisticación de las propias técnicas, como a su aplicación en nuevas áreas. Así, en el ámbito de la salud, se han aplicado para los procesos de diagnosis [RBD07], o en la clasificación de secuencias de ADN [GJLS11]. También se han aplicado en el sector industrial, para la detección de defectos en piezas de fundición [PLSSI$^+$12] o en el sector financiero, en los procesos de análisis del mercado de valores [HX99].

Por su parte, el sector natural donde aplicar este tipo de técnicas ha sido el propio sector TIC, en el que el recorrido del aprendizaje automático abarca cuestiones como la robótica y los juegos [dG05], el reconocimiento del habla y del lenguaje escrito [SW06] o el campo de los motores de búsqueda [JR07].

Las técnicas de aprendizaje automático se pueden clasificar en función del grado de interacción del hombre que necesitan. Aunque el propósito final del

AA es proporcionar conocimiento de modo automático, en ocasiones, para poder obtenerlo resulta necesaria la cooperación de las personas. En tal caso, la tarea suele consistir en que el ser humano aporte conocimiento del que dispone para que el sistema pueda seguir avanzando en la construcción de nuevo conocimiento que permita, finalmente, resolver la cuestión que pretende solucionar [Mit97]. En consecuencia, los tipos de aprendizaje según el grado de colaboración humana son:

- **Aprendizaje supervisado.** Este tipo de aprendizaje [RM98] consiste en un tipo de aprendizaje automático en donde al algoritmo que se usa para obtener el modelo relacional entre la entrada y las salidas, se le proporcionan una serie de ejemplos etiquetados que indican a qué clase pertenecen o el valor de salida, en el caso de problemas de regresión. Es decir, todos los ejemplos han sido clasificados a *priori*. De esta forma, en el proceso de aprendizaje, el algoritmo compara su salida actual con la etiqueta del ejemplo, para luego realizar los cambios que sean necesarios.

- **Aprendizaje no supervisado.** El aprendizaje no supervisado es un método de aprendizaje automático, en el que un modelo se ajusta a las observaciones dadas. La diferencia con el aprendizaje supervisado es que no existe un conocimiento previo. En el aprendizaje no supervisado no existe esa información que hace referencia a las salidas de los patrones que se van introduciendo, únicamente existen las entradas. Debe de ser el modelo, el cual vaya adecuando sus pesos dependiendo de la información interna que vaya recopilando de las entradas. Principalmente, esta clase de entrenamiento se emplea para clasificar y diferenciar rasgos significativos de un conjunto de datos no clasificado a *priori*, debido a que la red, internamente, intenta encontrar redundancias y rasgos significativos para agrupar los datos. Algunos algoritmos de agrupamiento o *clustering* como COBWEB [Fis87], EM [DLR77], o K-Medias [Llo82], son ejemplos de esta clase de aprendizaje.

En el desarrollo de esta investigación, hemos empleado el aprendizaje supervisado, ya que el conjunto de datos de mensajes y conversaciones obtenidos a partir de la red social *Facebook* fueron etiquetados previamente.

4.1.1 Las redes bayesianas

Las redes bayesianas son modelos gráficos que se emplean para representar las relaciones entre los sucesos, con el fin de inferir probabilidades e incertidumbres

asociadas a dichos sucesos, y por tanto estimar y predecir relaciones probables entre ellos. Estos modelos se basan en las teorías probabilísticas creadas en el siglo XVIII por *Thomas Bayes*. El uso de este tipo de modelos han resultado de utilidad en los algoritmos de búsqueda, dentro de una gran gama de aplicaciones que requieren recuperación de información, en el software de reconocimiento de imágenes y texto, entre otros.

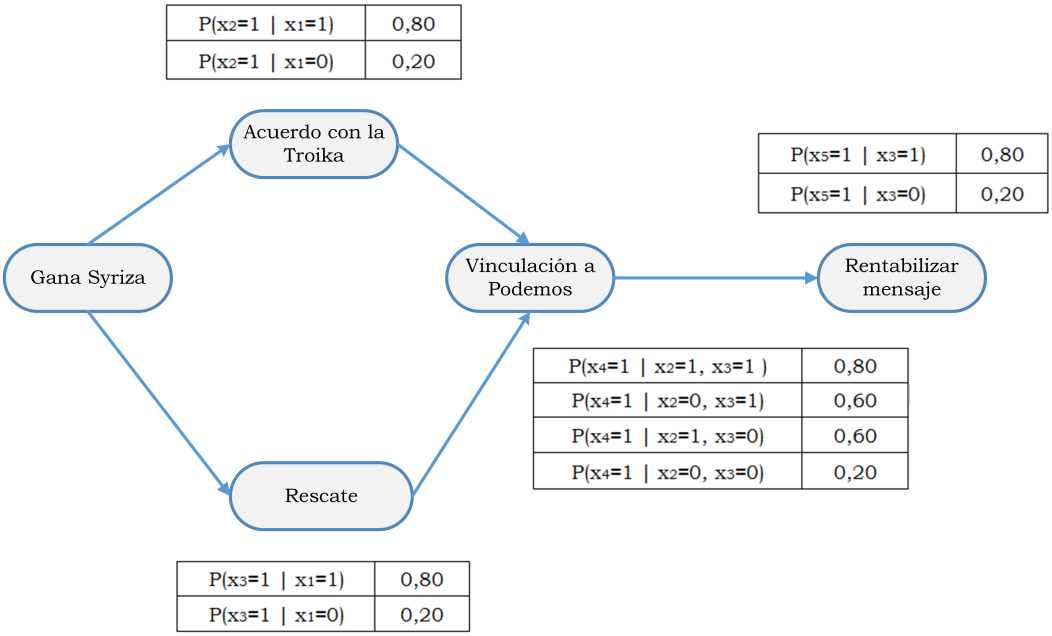


Figura 4.1: Ejemplo de clasificación mediante una red bayesiana.

Observando su formulación, se puede afirmar que dados dos eventos a los que denominaremos A y B, la probabilidad condicionada de que ocurra A habiéndose dado B, es decir $P(A|B)$, se puede obtener mediante la ecuación 4.1.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{4.1}$$

Gracias a este enfoque, se determinan las redes bayesianas (véase Figura 4.1), las cuales son modelos estadísticos basados en el teorema de Bayes [Pea82] y se definen como un modelo gráfico probabilístico para el aprendizaje y la predicción. Específicamente, las redes bayesianas se definen como un grafo acíclico dirigido[1] con una función de distribución probabilística asociada [CGH97]. Los

[1]Un grafo dirigido o digrafo es un tipo de grafo en el cual el conjunto de las aristas tiene una dirección definida, a diferencia del grafo generalizado, en el cual la dirección puede estar especificada o no.

nodos contenidos en el grafo dirigido ser refieren a las características de un problema concreto, mientras que las aristas representan las diferentes dependencias condicionales entre ellos. La función de probabilidad ilustra la robustez de las aristas en el grafo [CGH97].

Por su parte, existen diferentes algoritmos de aprendizaje basados en las redes bayesianas y que hemos empleado en esta investigación:

- **El algoritmo de aprendizaje K2.** El algoritmo K2 [CH91a] se encarga de buscar la red estructural con mayor probabilidad dado un conjunto de datos.
- **El clasificador *Naïve Bayes*.** Un clasificador bayesiano ingenuo [Lew98] asume la inexistencia de cualquier tipo de dependencia entre los nodos o atributos de la propia red bayesiana. Por lo cual, solamente se tienen que aprender las probabilidades de los valores de los atributos, dada la clase.
- **El método *Tree Augmented Naïve* (TAN).** El método TAN [FGG97] es un clasificador bayesiano simple aumentado con un árbol, el cual es una manera de mejorar la estructura del clasificador *Naïve Bayes* mediante la incorporación de ciertos nodos o atributos que tengan cierta dependencia.

4.1.2 Las máquinas de soporte vectorial

Las máquinas de soporte vectorial, del inglés *Support Vector Machines* o en su acrónimo SVM realizan una representación vectorial de los datos en un espacio n-dimensional. Así, dividen este espacio en 2 únicas regiones generando un hiperplano, que trata de maximizar la distancia o margen entre dichas 2 regiones. Este margen es la distancia más lejana entre las instancias de entrenamiento de ambas clases y se calcula a partir de la distancia entre las instancias de ambas clases al margen. Estas instancias se denominan *vectores soporte* [CV95] y sirven como referencia para clasificar el resto de instancias alrededor de las regiones establecidas, tal y como se observa en la Figura 4.2.

Formalmente, se define un conjunto de datos de entrenamiento $\mathscr{D}$, de la forma siguiente:

$$\mathscr{D} = \{(x_i, c_i) x_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^{n} \tag{4.2}$$

donde c_i denota la clase del punto x_i, situado en el rango (-1,1), y siendo a su vez, un vector de números reales en un espacio p-dimensional. El proósito de este algoritmo consiste en encontrar el hiperplano que maximice el margen m

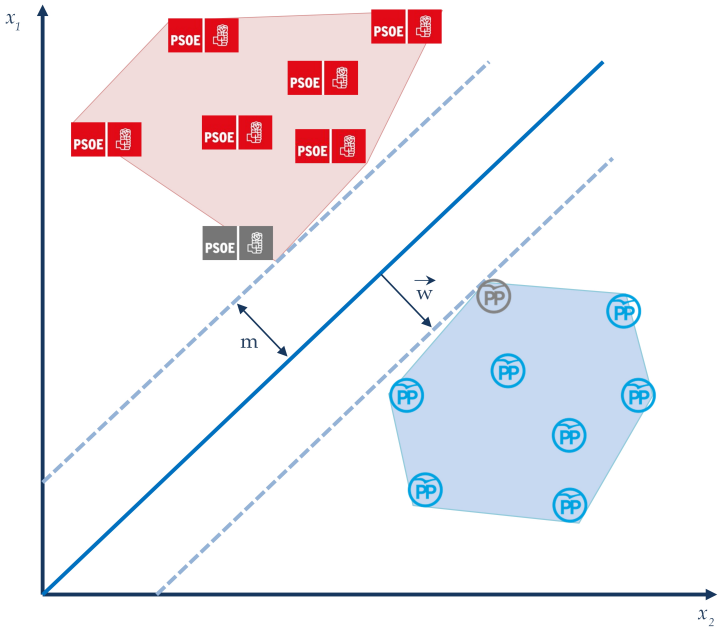


Figura 4.2: Ejemplo de clasificación mediante una máquina de soporte vectorial o *SVM*.

y que divida los puntos $c_i = 1$ de los puntos c_i = -1 en el espacio formado por el conjunto de datos $\mathscr{D}$. La representación del hiperplano se realiza mediante los puntos x_i, de modo que después de realizar el producto escalar entre x y w, y restado el valor de m, el valor resultante sea 0, siendo w un vector normal y ortogonal al hiperplano. Expresándolo de otro modo, todos los puntos x_i han de verificar la ecuación 4.3.

$$w \cdot x - m = 0 \tag{4.3}$$

De este modo se dispone de dos hiperplanos, correspondientes cada uno de ellos a cada clase de c_i, pudiéndose describir mediante la ecuación 4.4.

$$c_i(w \cdot x_i - m) \geq 1, \forall x_i 1 \leq i \leq n \tag{4.4}$$

Con mucha frecuencia, en lugar de utilizar hiperplanos lineales, se emplean funciones de núcleo que conducen a clasificaciones no lineales como polinomiales, radiales o sigmoidales [AW99]. Durante nuestras experimentaciones se ha hecho uso concreto de ciertas configuraciones del núcleo de las máquinas de soporte de vectorial que se enumeran a continuación:

- **El núcleo polinomial.** Este tipo de núcleo representa la similitud de los vectores (instancias de entrenamiento) en un espacio característico sobre los polinomios de las variables originales [AW99].

- **El núcleo polinomial normalizado.** El núcleo normalizado es una medida de la distancia entre 2 puntos de la superficie de una esfera controlado por un parámetro, análogo a la distancia *Mahalanobis*[1]. Esta interpretación es particularmente interesante cuando se usa con la proyección equidistante, en la cual la longitud del arco entre 2 puntos de la esfera es la misma que su separación en el plano medido por la distancia euclídea [WC00].

- **El núcleo basado en funciones de base radial.** Este núcleo mapea muestras de forma no lineal en un espacio dimensional superior; de modo que, a diferencia del núcleo lineal, es capaz de gestionar la instancia cuando la relación entre las etiquetas de las clases y los atributos es no lineal [HCL03]. Hay algunas situaciones donde el núcleo RBF no es adecuado. En particular, cuando el número de características es muy alto.

- **El núcleo *Pearson VII*.** Basado en la función *Pearson VII*[2], se caracteriza por su robustez y gran potencia de mapeo (comparada con los núcleos anteriores), con lo que puede ser usada como un núcleo universal capaz de ser una alternativa genérica a las funciones comunes de núcleo [ÜMB07].

4.1.3 Los *K*-vecinos más cercanos

El algoritmo *K*-vecinos más cercanos (*KNN*, del inglés *K-Nearest Neighbours*) [FH52] es un modelo de aprendizaje automático supervisado basado en instancias. Este método clasifica una instancia desconocida basándose en la clase de las instancias más cercanas a dicha instancia, en el espacio de entrenamiento. Para este fin, la instancia desconocida se coloca en este espacio y el algoritmo mide la distancia entre las instancias de entrenamiento y esta instancia desconocida. A pesar de los diferentes métodos existentes a la hora de elegir la clase de la instancia, la técnica más común es simplemente clasificarla como la clase más común entre los *K*-vecinos más cercanos.

[1] La distancia *Mahalanobis* es una forma de determinar la similitud entre 2 variables aleatorias multidimensionales. Se diferencia de la distancia euclídea en que considera la correlación entre las variables aleatorias.

[2] La función *Pearson VII* era una función popular en los años 1980 y 1990 para describir formas picudas a partir de patrones de difracción de rayos X convencionales.

A continuación desarrollamos el concepto de matemático del algoritmo K-vecinos más cercanos. En este clasificador basado en instancias, la fase de entrenamiento se inicia empleando un conjunto de instancias $\mathscr{S} = \{s_1, s_2, ..., s_{n-1}, s_n\}$, donde n representa el número de características de cada una de las instancias. Este conjunto de datos se representa en un espacio de entrenamiento como el mostrado en la Figura 4.3.

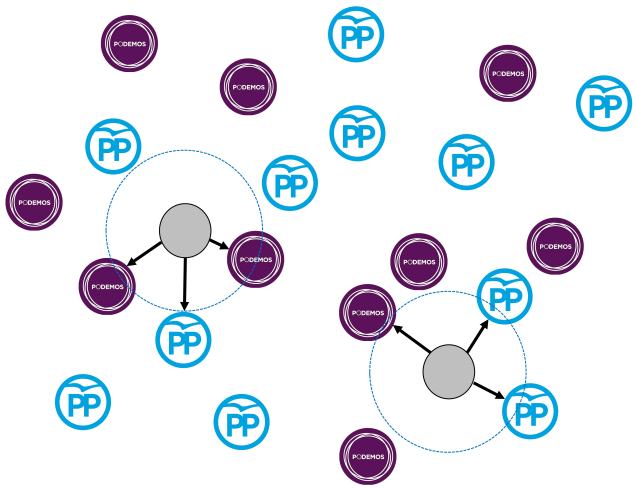


Figura 4.3: Ejemplo de clasificación mediante K-vecinos más cercanos.

Más adelante, y con la llegada de una nueva instancia a categorizar, comienza la fase de clasificación. Llegados a este punto, se calcula la distancia de la nueva instancia respecto a las instancias utilizadas en la fase de entrenamiento. Dicha distancia puede ser calculada mediante distintos tipos de métricas, como por ejemplo: (i) la distancia euclidiana[1], (ii) la distancia del coseno[2] o (iii) la distancia *Manhattan*[3], entre otras. En esta tesis hemos empleado la distancia euclidiana, cuya definición matemática es la mostrada en la ecuación 4.5.

$$\sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2} \tag{4.5}$$

Una vez calculadas las distancias, en base a la métrica elegida previamen-

[1]La distancia euclidiana o euclídea es la distancia *ordinaria* entre 2 puntos de un espacio euclidiano; es decir, la distancia que se mediría con una regla.

[2]La distancia del coseno es una medida de similitud entre 2 vectores de un espacio vectorial que mide el coseno del ángulo formado entre ellos.

[3]La distancia *Manhattan* es la distancia entre 2 puntos medida sobre los ejes a ángulos rectos; es decir, la distancia que se recorrería para llegar de un punto a otro si se siguiera una trayectoria cuadriculada.

te, podemos emplear distintos métodos para determinar la clase de la instancia desconocida. La técnica más empleada es elegir la clase más común entre las K instancias más cercanas pertenecientes al espacio de entrenamiento y la instancia a clasificar.

4.1.4 Los árboles de decisión

Los clasificadores de árboles de decisión (DT, de las voces inglesas *Decision Trees*) son un tipo de algoritmos pertenecientes al aprendizaje automático supervisado, los cuales son representados gráficamente en forma de árboles (véase la Figura 4.4). En este tipo de algoritmos, los nodos internos representan las condiciones de prueba con las variables de un problema concreto, mientras que los nodos finales u hojas representan la decisión final del algoritmo [Qui93].

Para conocer la estructura de árbol de estos modelos de un conjunto de datos etiquetado, existen diferentes métodos de entrenamiento a utilizar. En particular, hemos empleado los siguientes:

- **El algoritmo C4.5**. Este clasificador es una extensión del algoritmo ID3 [Qui86]. En la herramienta WEKA (acrónimo del inglés *Waikato Environment for Knowledge*), la implementación de este algoritmo se denomina J48 [Qui93]. Su funcionamiento se basa en un concepto denominado *entropía de la información*, la cual calcula la incertidumbre asociada a una variable aleatoria. En ocasiones, este término va asociado a la *entropía de Shannon* [Sha51], que mide el valor esperado de información dentro de un mensaje.

- **Los bosques aleatorios**. Otro método empleado son los bosques aleatorios (RF, del inglés *Random Forest*) [Bre01]. Este algoritmo genera varios árboles de decisión de forma aleatoria, combinándolos posteriormente como si fuesen un único clasificador. Concretamente, se genera un vector de entrada, con cada uno de estos árboles, formado por un subconjunto de datos elegidos mediante una discriminación estocástica [Kle96].

A continuación, desarrollamos el concepto matemático del algoritmo de árboles de decisión. Formalmente, se puede definir un árbol de decisión como un grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, donde $\mathscr{V}$ es un conjunto de nodos no vacío y $\mathscr{E}$ es un conjunto de aristas. Por lo cual, se define el camino del árbol como la secuencia de aristas que recorre, pudiendo ser representado de la forma $(v_1, v_2), (v_2, v_3), ..., (v_{n-1}, v_n)$. Por lo tanto, si tenemos en cuenta que (v, w) es una arista del árbol, se puede definir v como el nodo padre de w, y a su vez w como el nodo hijo de v. El único

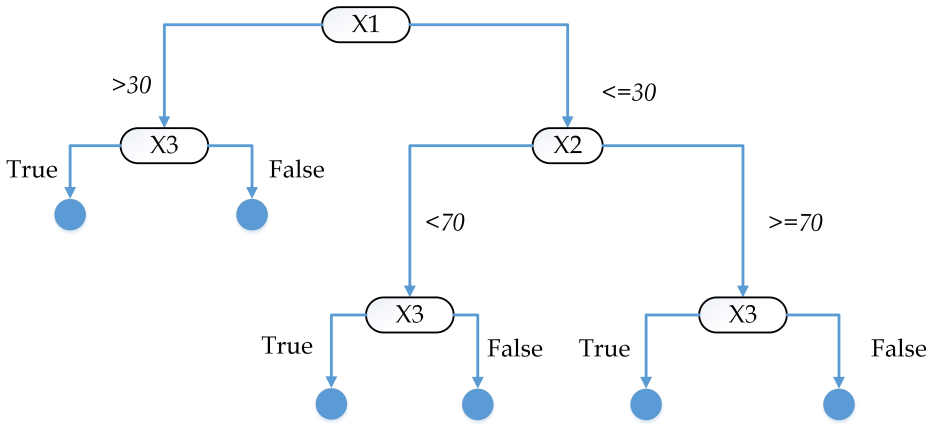


Figura 4.4: Ejemplo de clasificación mediante un árbol de decisión.

nodo en el árbol sin padres se denomina nodo raíz. Cualquier otro nodo en el árbol es un nodo interno. Para construir la representación gráfica de un árbol se contesta a un conjunto de preguntas binarias (sí o no).

Más adelante, cada uno de los árboles generados realiza una clasificación. Para determinar el resultado final, se emplea un sistema de votos en el que participan todas las clasificaciones de los árboles generados.

El principal inconveniente que presenta este clasificador es la sensibilidad al ruido que puedan presentar los nuevos datos [SLTW04]. Es decir, se adapta correctamente al conjunto de datos con los que realizamos el entrenamiento, pero pierde precisión en caso de que los futuros nuevos datos contengan ruido.

4.2 Validación empírica

Una vez descritas las técnicas de clasificación, procede describir la metodología que se ha llevado a cabo con el fin de validar los resultados obtenidos.

4.2.1 Metodología

Recordemos que los conjuntos de datos han sido pre-procesados con arreglo a sus respectivos modelos de datos, quedando recogida su información en ficheros ARFF. Para cada uno de estos conjuntos de datos -mensajes y conversaciones- se han definido las siguientes transformaciones:

- **SMOTE**. Como se ha descrito en la sección 3.4 del capítulo 3, los conjuntos de datos presentan un grado notable de dispersión y desbalanceo con

respecto a las clases definidas.

Con el uso de la técnica SMOTE es posible compensar esta circunstancia de los datos. Así, permite rebalancear las distribuciones de frecuencia [CBHK02], de modo que trata de sobrerepresentar las clases menos frecuentes al tiempo que infra-representa aquellas que cuentan con mayor frecuencia de aparición. de ahí su nombre: *Synthetic Minority Oversampling TEchnique*. Esta reasignación la hace creando muestras sintéticas a partir de los conjuntos de datos de partida y su asignación original de frecuencias.

El sobre-muestreo se realiza creando instancias sintéticas de la clase minoritaria de cada conjunto de datos de entrenamiento. La clase minoritaria es sobre-muestreada tomando cada muestra de la clase minoritaria e introduciendo instancias sintéticas a lo largo de los segmentos lineales uniendo con todos los k-vecinos más cercanos de la clase minoritaria. En función del grado de sobre-muestreo requerido, los vecinos pertenecientes a los k-vecinos más cercanos son elegidos aleatoriamente. De esta manera, las instancias seguirán siendo únicas y las clases pasarán a estar equilibradas.

- **Validación cruzada.** La validación cruzada (del inglés *cross validation*) es una práctica estadística, la cual permite dividir un conjunto de datos en subconjuntos de tal manera que el análisis se realiza inicialmente en uno de ellos, mientras que los subconjuntos restantes son retenidos para su posterior empleo en la confirmación y validación de dicho análisis inicial. Explicado de otra manera, se basa en la división del conjunto de datos en k subconjuntos distintos, utilizándose una parte de cada uno de los subconjuntos para entrenar el modelo y la otra para testearlo.

  En nuestros experimentos, hemos asignado el valor de 10 a la variable k, con lo que hemos dividido el conjunto de datos total en 10 subconjuntos diferentes. En cada uno de los subconjuntos hemos empleado el 90 % para entrenar el modelo y el 10 % restante para probarlo. Este método se aplica generalmente en la evaluación del aprendizaje automático [Bis95].

- **Aprendizaje del modelo.** Para cada subconjunto de datos, hemos efectuado la etapa de aprendizaje de cada algoritmo utilizando diferentes parámetros o distintos tipos de algoritmos de aprendizaje, dependiendo del modelo específico. Los algoritmos usan los parámetros por defecto en la herramienta de aprendizaje automático WEKA. En particular, hemos usado los siguiente modelos:

  1. *Redes bayesianas (BN ó Bayesian Networks)*.

Los algoritmos de aprendizaje estructural utilizados han sido las versiones K2 [CH91b] y *Tree Augmented Naïve* (TAN) [GGP$^+$97], además del clasificador *Naïve Bayes* [Bis95].

2. *Máquinas de soporte vectorial (SVM ó Support Vector Machines)*.
   Para este tipo de algoritmos se han utilizado el núcleo polinomial [AW99], un núcleo polinomial normalizado (del inglés *normalized polynomial kernel*) [MBM08], un núcleo *Pearson VII* [ÜMB07] y un núcleo basado en funciones de base radial (RBF ó *radial basis function*) [CYL$^+$08].
3. *K-vecinos más cercanos (KNN o K-nearest neighbour)*.
   Los experimentos realizados incluyen diferentes cantidades de vecinos (K), desde $K = 1$ hasta $K = 5$.
4. *Árboles de decisión (DT o Decision Trees)*.
   Para los experimentos, se han utilizado dos tipos de implementación. En primer lugar, se ha usado la implementación J48 que WEKA [G$^+$95] realiza del algoritmo *C4.5* [Qui93]) y en segundo lugar, *Random Forest*, un conjunto de árboles de decisión construidos al azar. Se han realizado experimentos con distintas cantidades de árboles aleatorios N, $N = 10$, $N = 50$, $N = 75$ y $N = 100$.

Evaluación del modelo Con el fin de validar empíricamente los resultados, resulta necesario establecer los indicadores correspondientes. Así, las métricas utilizadas son la precisión del modelo ((*ACC* del inglés *Precisión*) y el área bajo la curva ROC (AUC en su expresión anglosajona *Area Under the ROC Curve*). Ambos indicadores se basan en los conceptos de TPR y FPR.

- *Tasa de verdaderos positivos (TPR ó True Positive Rate)*. Representa la proporción de casos positivos correctamente clasificados entre el total de la población, y por consiguiente, la capacidad de clasificar correctamente los casos positivos. También se define como *sensibilidad*.

- *Tasa de falsos positivos (FPR, del inglés False Positive Rate)*. Representa la proporción de casos negativos correctamente clasificados entre el total de la población, y por consiguiente, la capacidad de clasificar correctamente los casos negativos. También se denomina *especificidad*.

$$TPR = \frac{TP}{TP + FN} \tag{4.6}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.7}$$

donde:

- TP es el número de verdaderos positivos (del inglés *true positive*), esto es, cuántos positivos han sido clasificados adecuadamente
- FN denota el número de falsos negativos (del inglés *false negative*), es decir, la cantidad de casos que han sido incorrectamente clasificados como tal.
- FP representa el número de falsos positivos (*false positive*) o positivos inadecuadamente clasificados.
- TN es el número de verdaderos negativos (*true negative*) ó negativos correctamente clasificados.

- *Precisión*. Es el número total de aciertos de los clasificadores dividido por el número total de casos en el conjunto de datos.
- *Área bajo la curva ROC*. También denominada AUC (*Area Under ROC curve*) es el área inferior delimitada por la curva definida en el *espacio ROC*, que es la zona definida del plano en la que se ponen en relación dos ejes cartesianos que representan la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos(FPR) [SKM09], esto es, la sensibilidad frente a $(1 - especificidad)$. De manera intuitiva, se puede interpretar como la probabilidad de que un clasificador ordenará una instancia positiva, elegida al azar, más alta que una instancia negativa.

4.3 Discusión de los resultados

En esta sección se presentan los resultados de los distintos procesos de experimentación. Tras haber etiquetado cada una de las instancias de mensajes y conversaciones en los conjuntos de datos, se ha procecido a la fase de entrenamiento de los mismos para, posteriormente, abordar la tarea de categorización. En esta tarea se muestra cómo las distintas técnicas son capaces de reconocer cada una los valores aprendidos.

En las páginas siguientes, se presentan los resultados ordenados de la manera siguiente. En primer lugar se exponen los resultados de la categorización de los mensajes y posteriormente los correspondientes a las conversaciones.

4.3.1 Conjunto de mensajes

Los mensajes se han categorizado en distintos tipos de clasificación que se expusieron en el capítulo 3, a saber: la función informativa, la función movilizadora, el nivel de meta-comunicación, el nivel de personalización, el grado de debate, la referencialidad de los contenidos, el nivel de campaña negativa, el macro primario y el marco secundario.

4.3.1.1 Función informativa

La aplicación de distintos clasificadores utilizados, tanto aplicando técnicas de *N-gramas* como *Tokens* o Palabras se recoge en la Tabla 4.1. En ella se comprueba que los mejores resultados son los proporcionados por los árboles de decisión (*AD*) usando bosques aleatorios N=50 es el mejor clasificador, ya que ofrece un compromiso entre precisión (97,21 %) y un área bajo la curva (*AUC*) suficientemente elevado (0,98).

En términos únicamente de precisión, los árboles de decisión con *Random Forest* aportan los resultados más altos, en todo caso superiores al 87 %, independientemente del número de árboles que se utilicen (en nuestro caso, 10, 25, 50, 75 y 100). Tanto las Redes Bayesianas (*RB*), en sus modalidades *K2* o *KAN*, obtienen resultados algo inferiores en términos de precisión, pero sustancialmente menores en lo que al *AUC* se refiere.

Por su parte, las *Supporting Vector Machines* o *SVM* ofrecen resultados desiguales en términos de precisión, en función de la variante que se utilice. Entre una familia u otra de *SVM* se produce una diferencia en la precisión de más de 24 puntos porcentuales. Finalmente, *Naïve Bayes* ofrece los resultados más pobres, tanto en lo referente a precisión como a *AUC*.

4.3.1.2 Función movilizadora

La Tabla 4.2 recoge los resultados de los distintos clasificadores utilizados, tanto aplicando técnicas de *N-gramas* como *Tokens*.

En términos generales, el uso de N-gramas otorga resultados superiores al uso de palabras para los mismos algoritmos en relación a la precisión: en 13 de ellos la preción de los N-gramas es superior, frente a 5 usos de *tokens*. En cambio, con respecto al valor del área bajo la curva ROC, emplear palabras consigue que 10 algoritmos obtengan mejores resultados que si usan N-gramas, que sólo superan a las palabras en un caso. Así mismo, 7 algoritmos obtienen el mismo resultado empleando cualquiera de los dos métodos.

Tabla 4.1: Resultados de la clasificación de la función informativa en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	96,65±0,75	0,89±0,02	96,65±0,73	0,90±0,03
RB: TAN	96,39±0,52	0,88±0,02	96,26±0,64	0,88±0,02
Naïve Bayes	16,13±1,11	0,58±0,02	12,61±0,95	0,58±0,02
SVM: Polinomial	96,45±0,69	0,97±0,02	96,45±0,64	0,98±0,01
SVM: Polinomial norm.	96,73±0,54	0,97±0,02	96,83±0,63	0,98±0,02
SVM: Pearson VII	96,70±0,56	0,97±0,02	96,67±0,65	0,97±0,02
SVM: RBF	96,32±0,55	0,98±0,01	96,86±0,58	0,98±0,01
KNN K=1	92,01±1,35	0,86±0,03	91,72±1,24	0,86±0,06
KNN K=2	93,76±0,91	0,91±0,01	93,54±0,88	0,92±0,03
KNN K=3	93,18±1,09	0,93±0,01	92,99±1,08	0,93±0,03
KNN K=4	93,94±0,95	0,94±0,01	93,91±1,08	0,94±0,02
KNN K=5	93,65±1,02	0,94±0,01	93,69±0,99	0,94±0,02
AD: J48	97,12±0,91	0,93±0,05	97,24±0,82	0,93±0,05
AD: Random Forest N=10	96,95±0,68	0,97±0,03	97,06±0,58	0,97±0,02
AD: Random Forest N=25	97,11±0,62	0,97±0,02	97,08±0,53	0,98±0,01
AD: Random Forest N=50	97,18±0,66	0,98±0,02	97,21±0,53	0,98±0,01
AD: Random Forest N=75	97,12±0,66	0,98±0,02	97,17±0,56	0,98±0,01
AD: Random Forest N=100	97,11±0,70	0,98±0,02	97,11±0,53	0,98±0,01

La mejor técnica de clasificación la ofrecen los árboles de decisión con 100 árboles y bosques aleatorios empleando N-gramas, que logran alcanzar una precisión del 89,06 % y un área bajola curva ROC de 0,79. De igual forma, los arboles de decisión con bosques aleatorios de más de 10 árboles, empleando indistintamente N-gramas o palabras, constituyen un grupo formado por 9 algoritmos con precisiones y valores de AUC en un entorno estadísticamente significativo similar al obtenido por el mejor algoritmo.

Las máquinas de soporte vecrtorial, los *KNN* y las redes bayesianas ocupan una zona central con distintas precisiones y valores de AUC. Finalmente, Naïve Bayes consigue los resultados más pobres, tanto en lo que se refiere a precisión (20,79 % y 14,27 %), como a área bajo la curva ROC, con un valor de 0,58.

Tabla 4.2: Resultados de la clasificación de la función movilizadora en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		Tokens	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	86,50±0,72	0,67±0,04	87,11±0,51	0,70±0,04
RB: TAN	86,19±0,79	0,64±0,03	86,34±0,90	0,65±0,03
Naïve Bayes	20,79±2,51	0,58±0,04	14,27±1,55	0,58±0,04
SVM: Polinomial	73,91±0,89	0,73±0,02	73,60±1,55	0,75±0,03
SVM: Polinomial norm.	86,79±0,84	0,75±0,02	85,60±0,80	0,76±0,03
SVM: Pearson VII	84,86±2,69	0,75±0,03	85,57±0,87	0,75±0,03
SVM: RBF	62,05±1,89	0,76±0,03	57,63±2,01	0,77±0,03
KNN K=1	64,03±1,57	0,62±0,03	61,39±2,48	0,62±0,02
KNN K=2	72,39±1,47	0,64±0,02	70,20±2,10	0,65±0,03
KNN K=3	68,77±1,91	0,65±0,03	65,99±2,58	0,67±0,03
KNN K=4	72,68±1,66	0,65±0,03	69,99±2,13	0,68±0,03
KNN K=5	72,10±2,18	0,65±0,03	68,81±2,53	0,68±0,03
AD: J48	81,27±1,73	0,65±0,04	81,71±2,03	0,63±0,05
AD: Random Forest N=10	87,55±0,96	0,73±0,05	87,61±0,51	0,74±0,04
AD: Random Forest N=25	88,75±0,64	0,77±0,04	88,55±0,76	0,77±0,04
AD: Random Forest N=50	88,91±0,63	0,78±0,04	88,59±0,77	0,78±0,04
AD: Random Forest N=75	89,01±0,50	0,78±0,04	88,69±0,63	0,79±0,03
AD: Random Forest N=100	89,06±0,51	0,79±0,03	88,66±0,41	0,79±0,04

4.3.1.3 **Grado de debate**

En la Tabla 4.3 se muestran los resultados de la categorización del grado de debate de los mensajes políticos.

En relación a la precisión del modelo, el uso de palabras ofrece resultados mejores que el de N-gramas en 11 algoritmos, frente a 7 de N-gramas. En términos de precisión, las palabras aportan mejores resultados en 8 algoritmos, mientras que los N-gramas ofrecen mejores resultados en 2 de ellos, siendo los resultados iguales usando uno u otro método.

La técnica que ofrece mejores resultados son los árboles de decisión con Random Forest y N=100 utilizando palabras, que logran una precisión del 60,67% y un área bajo la curva ROC con valor 0,78. Esa marca superior es compartida por 12 algoritmos en un entorno estadísticamente similar en términos de precisión y área bajo la curva ROC. Estos algoritmos son los árboles de decisión con bosques aleatorios de más de 10 árboles y las máquinas de soporte vectorial con núcleos Pearson VII y polinomial normalizado.

En el grupo intermedio se sitúan las redes bayesianas y los *K*-vecinos más próximos, con precisiones y valores de AUC heterogéneos. Por último, las redes bayesianas Naïve Bayes se sitúan en las posiciones con resultados más reducidos de precisión (entre 18,66 % y 23,1 % y AUC de 0,63 y 0,64 respectivamente), y en términos de área bajo la curva ROC, los *K*-vecinos más próximos ocupan las últimas posiciones, con valores de AUC iguales a 0,59 y 0,60, así como como precisiones de 39,88 % y 40,1 % respectivamente.

Tabla 4.3: Resultados de la clasificación del grado de debate en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Naïve Bayes	18,66±1,08	0,63±0,01	23,10±1,81	0,64±0,01
SVM: Polinomial	48,52±1,79	0,73±0,01	49,28±1,98	0,74±0,02
SVM: Polinomial norm.	58,47±2,17	0,76±0,01	59,11±1,42	0,77±0,01
SVM: Pearson VII	59,40±5,21	0,74±0,02	58,83±2,40	0,75±0,01
SVM: RBF	33,79±2,67	0,73±0,02	31,43±2,97	0,73±0,02
KNN K=1	39,88±2,01	0,59±0,02	40,10±1,97	0,60±0,02
KNN K=2	31,53±1,90	0,61±0,01	31,29±2,56	0,61±0,02
KNN K=3	30,08±1,90	0,62±0,01	29,13±2,43	0,62±0,02
KNN K=4	26,50±1,63	0,63±0,01	25,64±1,71	0,62±0,02
KNN K=5	25,68±1,89	0,63±0,01	24,24±1,46	0,62±0,01
AD: J48	48,69±2,14	0,65±0,01	49,70±1,72	0,66±0,02
AD: Random Forest N=10	52,78±1,33	0,72±0,02	52,91±2,37	0,72±0,02
AD: Random Forest N=25	57,27±1,68	0,75±0,01	57,34±1,77	0,75±0,02
AD: Random Forest N=50	59,55±1,67	0,77±0,01	59,90±1,35	0,77±0,02
AD: Random Forest N=75	60,16±1,53	0,77±0,01	60,18±1,44	0,77±0,02
AD: Random Forest N=100	60,56±1,34	0,78±0,01	60,67±1,62	0,78±0,02

4.3.1.4 **Grado de contribución**

Los resultados de clasificar el grado de contribución de un mensaje se presentan en la Tabla 4.4. Se puede observar que el reconocimiento y clasificación del grado de contribución de un mensaje resulta más eficaz utilizando árboles de decisión con bosques aleatorios de 75 árboles y utilizando N-gramas, alcanzando una precisión del 82,44 % y describiendo un área bajo la curva ROC de 0,83. Un poco por debajo estaría el uso de redes bayesianas -tanto usando N-gramas como palabras- que alcanzan el valor máximo de precisión de cualquiera de los algoritmos (82,34 %), aunque el área ROC es ligeramente inferior (0,81).

Los *K*-vecinos más próximos nos ofrecen los resultados menos favorables en su globalidad, ya sea utilizando cualquiera de los casos utilizados (con K tomando valores entre 1 y 5), o bien a través de cualquiera de las modalidades de palabras o N-gramas. Por su parte, utilizar máquinas de soporte vectorial, *Naïve Bayes* o el resto de técnicas otorga resultados intermedios, sin especial relevancia en términos de precisión o área bajo la curva ROC.

En la escala opuesta, las precisiones menores con las de *KNN* con K=2 utilizando palabras, que tan solo logra una precisión del 36,7 % y un AUC igual a 0,6. Así mismo, si utilizamos KNN con K=1 y palabras obtendremos el menor valor del área bajo la curva ROC, 0,57, con una precisión bastante limitada del 47,39 %.

En términos globales, el uso de N-gramas permite alcanzar mejores resultados de precisión (10 algoritmos frente a 8 que usan palabras), mientras que si se usan palabras se obtienen niveles de área bajo la curva ROC superiores en 13 de los 18 algoritmos utilizados.

Tabla 4.4: Resultados de la clasificación del grado de contribución en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	78,48±2,23	0,80±0,03	79,84±1,98	0,80±0,03
RB: TAN	82,34±1,76	0,81±0,02	82,34±1,30	0,81±0,02
Naïve Bayes	51,38±2,25	0,73±0,03	43,84±1,79	0,72±0,03
SVM: Polinomial	73,60±2,10	0,77±0,03	73,92±2,59	0,77±0,02
SVM: Polinomial norm.	80,18±1,96	0,78±0,03	79,84±1,88	0,78±0,03
SVM: Pearson VII	81,20±1,79	0,78±0,03	81,23±1,61	0,78±0,03
SVM: RBF	72,33±2,58	0,78±0,03	69,23±2,54	0,77±0,03
KNN K=1	47,89±1,79	0,58±0,02	47,39±1,83	0,57±0,01
KNN K=2	37,56±1,32	0,60±0,02	36,70±1,47	0,60±0,02
KNN K=3	45,03±1,92	0,62±0,02	44,81±1,68	0,62±0,02
KNN K=4	38,01±1,29	0,63±0,02	37,12±2,08	0,63±0,02
KNN K=5	43,71±1,69	0,64±0,02	42,46±1,67	0,65±0,02
AD: J48	76,51±1,77	0,72±0,03	76,16±2,05	0,72±0,03
AD: Random Forest N=10	78,99±2,39	0,79±0,03	79,56±2,12	0,79±0,03
AD: Random Forest N=25	81,74±2,11	0,81±0,02	81,90±1,89	0,81±0,02
AD: Random Forest N=50	81,89±2,07	0,82±0,03	82,12±2,14	0,82±0,02
AD: Random Forest N=75	82,24±1,85	0,83±0,03	82,28±2,06	0,82±0,02
AD: Random Forest N=100	82,32±1,73	0,83±0,03	82,24±2,15	0,82±0,02

4.3.1.5 **Nivel de meta-comunicación**

La Tabla 4.5 muestra los resultados de la categorización del nivel de meta-comunicación de los mensajes. En ella se aprecia que los mejores resultados para clasificar el nivel de meta-comunicación se obtiene mediante la técnica de N-gramas y con árboles de decisión (AD), y *Random Forest*. Si utilizamos 100 árboles se alcanza una precisión del 78,96 % y un *AUC* igual a 0,8. Para 75 árboles, la precisión se reduce mínimamente (78,93 %), manteniéndose el *AUC*.

Si en lugar de N-gramas utilizamos *tokens*, se conserva un *AUC* de 0,8 , pero la precisión desciende algunas décimas (78,79 % y 78,73 % para 100 y 75 árboles). En cambio, las redes *Naïve Bayes* continúan mostrando resultados muy pobres, no llegando ni al 20 % de precisión y con valores del área bajo la curva *ROC* inferiores a 0,53 en ambos casos.

Tabla 4.5: Resultados de la clasificación del nivel de meta-comunicación en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	75,56±0,72	0,69±0,02	75,72±0,82	0,70±0,02
RB: TAN	74,73±1,01	0,67±0,02	74,68±1,08	0,68±0,02
Naïve Bayes	16,10±1,29	0,53±0,02	10,30±1,04	0,52±0,02
SVM: Polinomial	66,49±1,40	0,77±0,01	65,70±1,39	0,77±0,01
SVM: Polinomial norm.	77,79±1,23	0,77±0,03	77,63±1,27	0,77±0,02
SVM: Pearson VII	78,16±1,08	0,79±0,02	77,09±1,14	0,79±0,02
SVM: RBF	58,70±1,56	0,79±0,02	56,32±1,79	0,79±0,01
KNN K=1	62,21±1,31	0,64±0,01	58,86±2,12	0,65±0,02
KNN K=2	68,74±0,97	0,66±0,01	65,93±1,71	0,67±0,02
KNN K=3	67,20±1,18	0,67±0,02	63,59±1,53	0,68±0,02
KNN K=4	69,26±0,96	0,68±0,02	65,98±1,48	0,69±0,02
KNN K=5	69,26±1,17	0,68±0,02	65,90±1,79	0,69±0,02
AD: J48	70,55±1,28	0,66±0,03	70,94±1,16	0,67±0,02
AD: Random Forest N=10	76,48±1,34	0,76±0,02	76,58±1,15	0,76±0,02
AD: Random Forest N=25	77,75±1,07	0,78±0,02	78,31±1,09	0,79±0,02
AD: Random Forest N=50	78,45±0,81	0,79±0,02	78,54±1,01	0,80±0,02
AD: Random Forest N=75	78,79±0,75	0,80±0,02	78,93±1,11	0,80±0,02
AD: Random Forest N=100	78,73±0,73	0,80±0,02	78,96±1,10	0,80±0,02

4.3.1.6 **Nivel de personalización**

La información completa sobre la clasificación del nivel de personalización de los mensajes se muestra en la Figura 4.6. Los resultados de N-gramas y Tokens son bastante similares cuantitativamente, aunque el uso de palabras hace que los algoritmos obtengan mejores resultados de precisión (en 14 algoritmos frente a 4 con N-gramas), e incluso mejores en relación con el área bajo la curva ROC (en 16 de los 18 algoritmos).

La técnica que provee el mejor resultado es el de los árboles de decisión con *Random Forest* utilizando 100 árboles, lo que alcanza un 79,33 % de precisión, y un área bajo la curva ROC igual a 0,87. Si se usan N-gramas, la precisión se reduce mínimamente, lográndose un 78,96 %, mientras que el *AUC* se mantiene en un 0,87.

El uso de *Naïve Bayes* queda relegado a las últimas posiciones: un 28,47 % de precisión y 0,64 como área bajo la curva ROC si se usan palabras, mientras que se la precisión se limita a un 26,25 % y el *AUC* toma como valor 0,65 cuando utiliza tokens.

El grupo de algoritmos *KNN* con $K = 1$ presentan los peores resultados de área bajo la curva ROC, con un *AUC* igual a 0,59 y precisión del 61,77 % si usa N-gramas, o un simple *AUC* de 0,6 y una precisión del 61,77 % si la opción es palabras.

4.3.1.7 **Referencialidad de contenidos**

La categorización de la referencialidad de los contenidos en los mensajes se muestra en la Figura 4.7. De la lectura de estos datos se puede afirmar que el uso de N-gramas es más eficaz con respecto a la precisión que puede lograr (10 de los 18 algoritmos obtienen una precisión más alta utilizando N-gramas, aunque con pequeñas diferencias). En cambio, para alcanzar un área bajo la curva ROC, el uso de palabras ofrece mejores valores en 13 de los 18 algoritmos, también con diferencias reducidas entre ambas técnicas.

Ahora bien, los árboles de decisión (*AD*) con bosques aleatorios de 75 árboles alcanzan los mejores resultados, tanto en precisión (91,99 %) como en AUC (0,87), seguidos a la par por los *AD* con bosques aleatorios de 50 árboles, que obtienen el mismo *AUC* (0,87) y una precisión de 91,98 %. En general, los árboles de decisión con *Random Forest* a partir de 25 árboles ofrecen los mejores resultados en precisión. Como técnicas menos efectivas, el uso de *Naïve Bayes*, con palabras y con N-gramas proporcionan eficacias y áreas bajo la curva ROC los peores resultados: precisión del 34,6 % y *AUC* igual a 0,75.

Tabla 4.6: Resultados de la clasificación del nivel de personalización en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	65,89±1,14	0,78±0,02	65,49±2,20	0,78±0,02
RB: TAN	67,67±1,82	0,80±0,02	67,44±2,50	0,81±0,02
Naïve Bayes	26,25±2,50	0,65±0,02	28,47±1,69	0,64±0,02
SVM: Polinomial	66,60±1,90	0,81±0,01	68,15±1,95	0,81±0,02
SVM: Polinomial norm.	74,40±1,47	0,84±0,01	76,36±1,71	0,86±0,01
SVM: Pearson VII	72,52±2,00	0,81±0,01	75,25±2,15	0,83±0,01
SVM: RBF	51,57±3,83	0,77±0,02	49,99±4,52	0,78±0,02
KNN K=1	61,77±2,48	0,60±0,01	62,51±2,50	0,59±0,02
KNN K=2	61,32±2,39	0,61±0,02	61,64±2,20	0,61±0,02
KNN K=3	59,55±3,10	0,61±0,01	60,63±2,91	0,61±0,02
KNN K=4	60,44±3,24	0,62±0,01	62,45±3,05	0,62±0,02
KNN K=5	59,71±3,38	0,62±0,01	62,54±3,25	0,62±0,02
AD: J48	72,05±2,11	0,77±0,02	72,56±1,76	0,78±0,01
AD: Random Forest N=10	72,39±2,51	0,81±0,02	71,92±2,21	0,81±0,02
AD: Random Forest N=25	76,35±1,74	0,85±0,01	76,36±1,31	0,85±0,01
AD: Random Forest N=50	77,85±1,65	0,86±0,01	78,17±0,89	0,86±0,01
AD: Random Forest N=75	78,85±1,69	0,87±0,01	78,85±1,10	0,87±0,01
AD: Random Forest N=100	78,96±1,66	0,87±0,01	79,33±1,00	0,87±0,01

4.3.1.8 **Nivel de campaña negativa**

El análisis de las técnicas para identificar el nivel de campaña que se explicita en los mensajes políticos queda recogido en la Tabla 4.8.

En este caso, los árboles de decisión (*AD*) usando bosques aleatorios de 100 árboles proveen los mejores resultados cuando se usan N-gramas. En ese caso se obtiene la mayor precisión (89,96%) y con *AUC* de 0.87. Si reducimos el número de árboles, se comprueba que se va cediendo mínimamente precisión. Se constata que, en todo caso, los árboles de decisión con *Random Forest* usando más de 50 árboles otorga una precisión y *AUC* altamente satisfactorios.

Los resultados de *SVM*, con *Pearson VII*, tanto usando palabras como N-gramas también son muy notables, logrando una precisión por encima del 89%, aunque con un AUC ligeramente inferior a los árboles de decisión con *Random Forest*, que logran un AUC de 0,84. De manera similar, cuando se trata de *SVM* con el núcleo polinomial normalizado, los resultados son un poco inferiores (por encima del 89% de precisión y *AUC* igual superior a 0,84).

Tabla 4.7: Resultados de la clasificación de la referencialidad de contenidos en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	83,62±2,44	0,81±0,04	84,16±1,76	0,82±0,04
RB: TAN	87,36±1,41	0,81±0,04	87,31±2,12	0,83±0,05
Naïve Bayes	33,78±2,88	0,73±0,03	34,60±2,96	0,75±0,02
SVM: Polinomial	75,38±1,64	0,80±0,01	75,66±2,03	0,80±0,03
SVM: Polinomial norm.	87,67±0,98	0,82±0,02	87,33±1,34	0,82±0,02
SVM: Pearson VII	89,48±1,21	0,78±0,06	89,91±0,94	0,82±0,03
SVM: RBF	55,11±1,82	0,84±0,02	52,01±2,04	0,84±0,02
KNN K=1	81,58±1,24	0,76±0,02	80,51±1,94	0,75±0,03
KNN K=2	74,89±2,73	0,78±0,02	72,97±3,28	0,76±0,03
KNN K=3	68,93±3,34	0,78±0,02	65,93±3,64	0,77±0,03
KNN K=4	63,75±3,34	0,78±0,02	60,10±4,36	0,77±0,03
KNN K=5	59,53±4,06	0,79±0,03	54,68±5,35	0,78±0,03
AD: J48	83,26±1,18	0,81±0,02	85,57±1,37	0,83±0,02
AD: Random Forest N=10	89,25±0,60	0,84±0,02	88,93±0,87	0,84±0,02
AD: Random Forest N=25	91,57±0,81	0,86±0,02	91,26±0,73	0,86±0,02
AD: Random Forest N=50	91,80±0,69	0,87±0,02	91,98±0,52	0,87±0,02
AD: Random Forest N=75	91,94±0,76	0,87±0,02	91,99±0,70	0,87±0,02
AD: Random Forest N=100	91,89±0,70	0,87±0,02	91,98±0,63	0,87±0,02

Las redes bayesianas, *RB*, pese a que alcanzan una buena precisión (no inferiores al 85 %), no se pueden considerar suficientemente positivos en términos del *AUC* (entre 0,65 y 0.72). Otro tanto sucede con *KNN*, mediante los que se tienen resultados con una precición moderadamente alta, pero bastante inferiores en cuanto a *AUC*. Por otro lado, y siguiendo la tónica del resto de clasificaciones, las técnicas de *Naïve Bayes* sigue proporcionando los resultados más pobres, a gran distancia del resto, tanto para palabras como a través del uso de N-gramas.

4.3.1.9 **Marco primario**

La clasificación del marco primario de los mensajes arroja unos resultados que se muestran en la Tabla 4.9. De ellos, se observa que a la hora de obtener mejor precisión, el uso de N-gramas ha resultado más eficaz en 15 de los 18 algoritmos, mientras que el el uso de tokens ha provisto mejores valores del área bajo la curva ROC en todos los casos menos uno.

Tabla 4.8: Resultados de la clasificación del nivel de campaña negativa en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	85,77±0,55	0,66±0,03	86,18±0,58	0,72±0,04
RB: TAN	86,85±0,40	0,65±0,03	86,49±0,72	0,67±0,03
Naïve Bayes	18,12±0,75	0,55±0,04	10,09±0,74	0,52±0,04
SVM: Polinomial	82,22±1,79	0,82±0,03	81,74±1,30	0,82±0,02
SVM: Polinomial norm.	89,55±0,88	0,84±0,03	89,10±0,81	0,85±0,02
SVM: Pearson VII	89,67±0,85	0,84±0,02	89,44±0,64	0,84±0,02
SVM: RBF	82,69±1,52	0,85±0,02	81,49±1,41	0,85±0,02
KNN K=1	76,26±1,57	0,63±0,02	74,22±1,35	0,64±0,01
KNN K=2	82,45±0,86	0,64±0,03	80,23±1,30	0,65±0,02
KNN K=3	80,38±0,86	0,64±0,03	77,93±1,71	0,65±0,02
KNN K=4	82,72±1,30	0,64±0,03	80,31±1,59	0,66±0,02
KNN K=5	82,13±1,10	0,64±0,03	80,04±1,54	0,66±0,01
AD: J48	87,22±0,59	0,72±0,03	88,11±1,15	0,76±0,03
AD: Random Forest N=10	88,75±1,20	0,83±0,02	87,98±1,07	0,82±0,02
AD: Random Forest N=25	89,41±0,72	0,86±0,02	88,79±0,80	0,85±0,02
AD: Random Forest N=50	89,64±0,79	0,87±0,02	89,25±0,73	0,86±0,02
AD: Random Forest N=75	89,74±0,88	0,87±0,02	89,38±0,82	0,87±0,02
AD: Random Forest N=100	89,96±0,81	0,87±0,02	89,29±0,93	0,87±0,02

En relación a los mejores resultados en precisión y área bajo la curva ROC, los árboles de decisión con 100 árboles en bosques aleatorios proporcionan las mejores cifras, que son sensiblemente inferiores a la mayoría de las clasificaciones hasta el momento. Así, la precisión máxima alcanzada se limita a un 52,61 % y un 0,8 como valor del área bajo la curva ROC, tanto si se emplean N-gramas como palabras. En la misma dirección, las variantes utlizadas de la familia de algoritmos *K*-vecinos más próximos, así como las redes bayesianas en sus núcleos *K2* y *TAN* no ofrecen diferencias estadísticamente significativas en relación a la precisión alcanzada.

En cambio, *Naïve Bayes* ofrece la precisión menor de todos los modelos, apenas un 19,02 % y un área bajo la curva ROC de 0,66 si se utilizan palabras, o del 18,25 % y 0,67 respectivamente si se utilizan N-gramas. Otro tanto sucede con *K*-vecinos más próximos con K=1, que produce los menores valores del área bajo la curva ROC (0,56) tanto cuando se emplean N-gramas (con precisión del 39,98 %) como con palabras (con una precisión del 38,79 %).

Tabla 4.9: Resultados de la clasificación del marco primario en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	49,16±1,19	0,76±0,01	48,82±1,39	0,77±0,01
RB: TAN	49,07±2,34	0,78±0,01	47,77±2,50	0,78±0,01
Naïve Bayes	18,25±1,40	0,67±0,01	19,02±1,59	0,66±0,01
SVM: Polinomial	41,21±1,27	0,75±0,01	41,99±1,68	0,75±0,01
SVM: Polinomial norm.	46,34±1,41	0,78±0,01	46,46±2,69	0,78±0,01
SVM: Pearson VII	33,59±6,24	0,73±0,02	35,00±6,11	0,74±0,02
SVM: RBF	31,07±1,85	0,73±0,01	26,54±2,35	0,71±0,01
KNN K=1	39,98±2,59	0,56±0,02	38,79±1,67	0,56±0,01
KNN K=2	37,43±2,01	0,58±0,02	35,68±2,22	0,57±0,02
KNN K=3	35,09±2,09	0,58±0,02	32,32±2,85	0,57±0,02
KNN K=4	34,04±3,01	0,58±0,02	31,25±3,14	0,57±0,02
KNN K=5	33,31±3,18	0,59±0,02	29,94±3,33	0,58±0,02
AD: J48	37,53±2,07	0,62±0,02	39,43±1,94	0,64±0,01
AD: Random Forest N=10	45,32±1,93	0,73±0,01	45,30±1,85	0,73±0,01
AD: Random Forest N=25	49,79±2,58	0,77±0,02	49,80±1,83	0,77±0,01
AD: Random Forest N=50	51,94±1,84	0,79±0,01	51,35±1,94	0,79±0,01
AD: Random Forest N=75	52,11±1,51	0,79±0,01	52,32±1,65	0,79±0,01
AD: Random Forest N=100	52,61±1,77	0,80±0,01	52,61±2,02	0,80±0,01

4.3.1.10 **Marco secundario**

La clasificación del marco secundario presenta los resultados que se muestran en la Tabla 4.10. El uso de palabras permite obtener mejores precisiones, aunque presenta heterogeneidad en los resultados. En 11 de los algoritmos obtiene mejores resultados, frente a otros 7 casos en los que obtiene valores inferiores. Por otro lado, el empleo de N-gramas permite que 9 algoritmos obtengan mayores valores de *AUC*, mientras que 3 obtienen los mismos resultados que utilizando palabras, y 6 aumentan su área bajo la curva ROC si usan palabras.

Los mejores resultados son los que alcanzan los árboles de decisión empleando N-gramas con bosques aleatorios de 100 árboles, que logran una precisión del 89,31 % y un área bajo la curva ROC de 0,81. En un entorno similar se encuentran el resto de árboles de decisión similares que emplean N-gramas con bosques aleatorios con menos árboles (N=25, 50, 75), con precisiones no inferiores al 88,05 % y *AUC* por encima de 0,78.

Los algoritmos K-vecinos más próximos con K=1 y K=2, ambos mediante palabras, logran precisiones ligeramente superiores (90,43 % y 89,57 % respec-

tivamente), pero el área bajo la curva ROC es muy inferior, tan solo 0,52 y 0,53.

El algoritmo que proporciona la precisión má pobre es *Naïve Bayes*, con un 14,46 % y *AUC* igual a 0,62 usando palabras, y 11,38 % y un área bajo la curva ROC de 0,58 mediante N-gramas.

Tabla 4.10: Resultados de la clasificación del marco secundario en mensajes utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec(%)**	**AUC**
RB: K2	87,83± 1,41	0,75±0,05	87,41± 1,09	0,77±0,04
RB: TAN	75,19± 2,94	0,77±0,04	75,09± 2,84	0,77±0,04
Naïve Bayes	11,38± 1,46	0,58±0,03	14,46± 1,42	0,62±0,04
SVM: Polinomial	67,71± 5,41	0,73±0,04	76,52± 3,42	0,71±0,04
SVM: Polinomial norm.	83,89± 3,55	0,76±0,03	84,50± 3,74	0,75±0,03
SVM: Pearson VII	80,59±27,25	0,66±0,10	84,87±23,10	0,64±0,10
SVM: RBF	64,85± 2,92	0,79±0,03	58,95± 3,82	0,77±0,04
KNN K=1	87,64± 2,17	0,56±0,04	90,43± 0,62	0,52±0,03
KNN K=2	83,87± 3,69	0,58±0,05	89,57± 1,01	0,53±0,03
KNN K=3	80,50± 4,59	0,57±0,06	88,77± 1,32	0,53±0,03
KNN K=4	77,50± 5,19	0,57±0,05	87,28± 1,81	0,54±0,03
KNN K=5	75,15± 5,61	0,57±0,04	86,18± 2,09	0,55±0,03
AD: J48	63,26± 3,69	0,61±0,04	64,70± 2,87	0,65±0,03
AD: Random Forest N=10	85,64± 1,72	0,76±0,04	85,24± 1,05	0,77±0,04
AD: Random Forest N=25	88,05± 1,48	0,78±0,04	87,47± 1,13	0,79±0,04
AD: Random Forest N=50	88,96± 1,14	0,80±0,03	88,53± 1,15	0,80±0,03
AD: Random Forest N=75	89,29± 1,07	0,80±0,03	89,23± 1,00	0,81±0,03
AD: Random Forest N=100	89,31± 1,02	0,81±0,04	89,44± 0,92	0,81±0,03

4.3.1.11 **Resumen**

En síntesis, los resultados de categorización de los mensajes expuestos hasta el momento muestran la capacidad de identificación y clasificación efectiva de mensajes publicados por ciudadanos en los muros digitales que los partidos políticos tienen en sitios de interacción social.

A fin de ser precisos, esta capacidad de clasificación de los mensajes es notable en el caso de la función informativa, el grado de referencialidad de los contenidos, el nivel de campaña negativa, el marco secundario, y la función movilizadora. Todos ellos ofrecen precisión en el rango del 90 %-97 %. De igual manera, el modelo es capaz de asignar categoría con gran capacidad en el grado

de contribución, el nivel de personalización y el nivel de meta-comunicación, ofreciendo en conjunto precisiones superiores en la franja del 80 %-90 %. En contraste con esto, el modelo encuentra más dificultades a la hora de clasificar el marco primario en los mensajes (precisión ligeramente superior al 50 %). Para más detalle, véase la Tabla 4.11.

Para alcanzar tales prestaciones que permiten clasificad cada una de dichas dimensiones de las conversaciones, la técnica más eficaz han sido los árboles de decisión con bosques aleatorios, preferentemente con un número de árboles a partir de 50. En cuanto al tipo de tratamiento del texto -mediante N-gramas o palabras- no se aprecian diferencias significativas respecto a su eficacia, motivo por el cual no es posible formular una recomendación de uso a tal respecto.

Dados estos niveles de clasificación, cabe señalar que la utilización de estas clasificaciones de mensajes pueden ser utilizadas con fiabilidad en la clasificación de conversaciones.

Tabla 4.11: Cuadro resumen con las mejores técnicas para categorizar mensajes, en términos de precisión (%) y área bajo la curva ROC (*AUC*).

Clasificación	**Técnica**	**Prec.(%)**	**AUC**
Función informativa	*AD: Random Forest N=50 (palabras)*	97,21	0,98
Función movilizadora	*AD: Random Forest N=100 (N-gramas)*	89,06	0,79
Grado contribución	*AD: Random Forest N=100 (N-gramas)*	82,32	0,83
Meta-comunicación	*AD: Random Forest N=100 ()palabras)*	78,96	0,8
Grado de referencialidad	AD: Random Forest N=75 ()palabras)	91,99	0,87
Nivel de personalización	*AD: Random Forest N=100 (palabras)*	79,33	0,87
Marco primario	*AD: Random Forest N=100 (N-gramas)*	52,61	0,8
Marco secundario	*AD: Random Forest N=100 (palabras)*	89,44	0,81
Campaña negativa	*AD: Random Forest N=100 (N-gramas)*	89,96	0,87

4.3.2 **Conjunto de conversaciones**

Los conversaciones políticas se han categorizado en distintos tipos de clasificación, a saber: la función informativa, la función movilizadora, el nivel de meta-comunicación, el nivel de personalización, la referencialidad de los contenidos, el nivel de campaña negativa, el macro primario y el marco secundario.

4.3.2.1 **Función informativa**

A la hora de categorizar la función informativa de las conversaciones (Véase la Tabla 4.12), los resultados alcanzados muestran que el uso de N-gramas ofrece mejores resultados que el de *palabras*, tanto a la hora de medir la precisión (en 10 de las técnicas el uso de N-gramas da mejores resultados frente a 8 usando palabras) como en lo que respecta a la precisión. Cuando se utilizan N-gramas, 13 de las técnicas consiguen mejores resultados de AUC que cuando se usan palabras, que son mejores en 5 técnicas.

Los mejores resultados de categorización son los alcanzados por las máquinas de soporte vectorial utilizando el núcleo Polinomial normalizado con *palabras*, con una precisión del 68,57 % y un área bajo la curva ROC de 0,73. En segundo lugar, SVM con el núcleo polinomial utilizando N-gramas logra una precisión ligeramente inferior (67,32 %), aunque con una AUC mejor, 0,77.

En sentido contrario, el peor rendimiento en términos de eficacia es la que se consigue usando *KNN* con palabras y K=5, que tan solo llega al 43,93 % y un AUC de 0,73, seguido de los árboles de decisión con el algoritmo J48 utilizando *palabras*, que tan solo logra una eficacia del 42,86 % y un 0,6 en el área bajo la curva ROC.

Con respecto al AUC, los árboles de decisión ofrecen los mejores resultados. Cuando se utiliza *Random Forest* con 100 árboles y *palabras* o bien al utilizar N-gramas y Random Forest con 50 árboles se consigue un área bajo la curva ROC de 0,79. El AUC es ligeramente inferior (0,78) con SVM con el núcleo polinomial y usando *Palabras*.

El uso de árboles de decisión mediante la implementación *J48* utilizando palabras, y los *KNN* (con K=1) nos muestran las cifras más pobres en AUC (0,6 y 0.59 respectivamente), y también un grado de eficacia bastante limitado (42,86 % y 44,82 % respectivamente).

4.3.2.2 **Función movilizadora**

La clasificación de la función movilizadora de las conversaciones ofrece los resultados que se muestran en la Tabla 4.13. El mejor resultado es el que proporciona la técnica SVM mediante el núcleo Polinomial y usando Palabras, que logra los mejores datos, tanto en lo que se refiere a precisión (alcanza el 71,07 %) como en el área bajo la curva ROC (0,79). En general, el uso de N-gramas mejora los resultados a los obtenidos mediante Palabras en 16 de las 18 técnicas utilizadas, tanto en eficacia como en área bajo la curva ROC.

De manera pormenorizada, la precisión alcanzada es mejor para las técnicas

Tabla 4.12: Resultados de la clasificación de la función informativa en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	58,39±16,37	0,78±0,15	58,39±17,40	0,73±0,14
RB: TAN	63,57±15,80	0,76±0,13	54,82±21,11	0,77±0,16
Naïve Bayes	59,46±14,19	0,70±0,13	58,39±15,27	0,71±0,12
SVM: Polinomial	67,32±11,76	0,77±0,09	62,14±11,41	0,78±0,09
SVM: Polinomial norm.	61,07±15,77	0,72±0,13	68,57±11,85	0,73±0,12
SVM: Pearson VII	48,93±22,86	0,69±0,14	63,57±13,42	0,69±0,16
SVM: RBF	62,32±17,40	0,68±0,15	56,07±18,54	0,72±0,17
KNN K=1	44,82±22,61	0,59±0,13	54,64±20,22	0,68±0,15
KNN K=2	54,64±20,22	0,65±0,09	55,89±19,42	0,73±0,16
KNN K=3	55,71±17,79	0,64±0,09	51,96±18,20	0,73±0,17
KNN K=4	49,64±13,78	0,62±0,11	51,96±18,20	0,74±0,12
KNN K=5	43,39±19,16	0,63±0,15	43,93±17,58	0,73±0,11
AD: J48	55,18±15,28	0,67±0,14	42,86±15,41	0,60±0,13
AD: Random Forest N=10	65,00±17,48	0,75±0,18	54,82±16,50	0,68±0,10
AD: Random Forest N=25	59,82±19,09	0,74±0,20	54,82±14,24	0,73±0,09
AD: Random Forest N=50	59,82±17,17	0,79±0,13	62,32±14,10	0,77±0,08
AD: Random Forest N=75	57,32±13,00	0,76±0,13	59,82±13,81	0,78±0,09
AD: Random Forest N=100	58,57±14,11	0,77±0,13	62,32±15,28	0,79±0,09

Naïve Bayes mediante el uso de *Palabras* (un 74,46%), con un AUC de 0,74, seguida del SVM con el núcleo Polinomial y utilizando *Palabras*, que obtiene un 71,07% y un AUC algo mejor (0,79). Por contra, KNN con K=2 y K=4 usando N-gramas consigue los resultados más pobres en eficacia, superando ligeramente el 20% (23,75% y 23,75% respectivamente) y con valores de AUC de 0,52 y 0,5.

El uso de *Palabras* vuelve a proporcionar los mejores resultados en el caso de SVM con el núcleo Polinomial ya comentado anteriormente, y con el algoritmo SVM y el núcleo Polinomial normalizado se obtiene una eficacia del 68,57% y un AUC con valor de 0,78. De nuevo, KNN otorga los peores resultados en área bajo la curva ROC para K=3 y K=4 (con valores para AUC de 0,46 y 0,5 respectivamente). Así mismo, la precisión es igualmente insuficiente (26,25% y 23,75%).

Tabla 4.13: Resultados de la clasificación de la función movilizadora en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	59,46± 7,90	0,60±0,11	61,07±16,83	0,61±0,17
RB: TAN	64,82±13,88	0,62±0,16	68,57±11,85	0,71±0,13
Naïve Bayes	68,75±13,50	0,57±0,17	74,46±12,52	0,74±0,14
SVM: Polinomial	68,57± 8,43	0,72±0,14	71,07±11,60	0,79±0,19
SVM: Polinomial norm.	64,64± 4,60	0,70±0,16	68,57±10,28	0,78±0,17
SVM: Pearson VII	45,54±26,37	0,56±0,25	58,21±15,59	0,62±0,25
SVM: RBF	55,54±19,26	0,66±0,21	69,64± 6,24	0,72±0,14
KNN K=1	31,61±15,75	0,50±0,11	31,96±16,93	0,51±0,11
KNN K=2	23,75±13,76	0,52±0,10	39,46±14,57	0,58±0,10
KNN K=3	26,25±22,40	0,46±0,12	39,46±16,78	0,62±0,12
KNN K=4	23,75±19,05	0,50±0,13	36,96±21,43	0,64±0,15
KNN K=5	24,11±13,84	0,51±0,14	35,36±17,29	0,64±0,12
AD: J48	51,96±17,22	0,52±0,17	55,71±16,79	0,53±0,22
AD: Random Forest N=10	68,21± 9,40	0,64±0,20	60,71±12,49	0,54±0,21
AD: Random Forest N=25	65,71± 6,62	0,66±0,18	64,46±10,15	0,66±0,22
AD: Random Forest N=50	64,46± 5,80	0,65±0,15	67,32±20,40	0,73±0,20
AD: Random Forest N=75	64,46± 5,80	0,70±0,15	68,57±15,64	0,71±0,21
AD: Random Forest N=100	63,21± 4,47	0,71±0,16	69,82±15,60	0,70±0,20

4.3.2.3 **Nivel de meta- comunicación**

De los resultados de la clasificación de conversaciones (véase Tabla 4.14 se determina que el mejor clasificador es el correspondiente a los árboles de decisión (AD) con *palabras* usando bosques aleatorios con 10 árboles ($N = 10$), que proporciona una precisión del 48,21 % y un área bajo la curva ROC de 0,66.

A la vista de los resultados anteriores, el uso de N-gramas es más favorable que *Palabras*, ya que la precisión obtenida es superior en 17 algoritmos, mientras que utilizar Tokens solo da mejores resultados en 1 algoritmo. En términos de AUC, N-gramas ofrece mejor resultado en 14 algoritmos, frente a 4 en los que las palabras dan mayor área bajo la curva.

Analizando los resultados para cada una de las métricas definidas, en términos de precisión, las mejores cifras son las alcanzadas por los árboles de decisión comentados (*Random Forest* con N=10), y en segunda posición el uso de máquinas de soporte vectorial con el núcleo polinomial normalizado, que logra una precisión del 45,89 % y un valor de AUC igual a 0,64 utilizando tokens.

En el otro extremo, los peores resultados en precisión son los obtenidos por los K-vecinos más próximos con N-gramas, con K=3 y K=5, que apenas consiguen una precisión superior al diez por ciento (12,68 % y 10 % respectivamente) y un AUC de 0,57 y 0,54 igualmente.

En relación al área bajo la curva ROC, los árboles de decisión usando palabras con bosques aleatorios (con 50 y 25 árboles) alcanzan los valores más altos, con 0,68 y 0,67 respectivamente (y unas precisiones del 44,46 % y 41,96 %). Sin embargo, en el lado opuesto, los K-vecinos más próximos usando palabras obtienen sendos valores de 0,51 en AUC (para $K = 1$ y $K = 2$), a la vez que muestran una eficacia muy pobre (27,86 % y 25,36 % respectivamente)

Tabla 4.14: Resultados de la clasificación del nivel de meta-comunicación en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	38,04±19,62	0,60±0,19	43,21±22,58	0,62±0,18
RB: TAN	39,29±13,81	0,61±0,16	44,46±19,26	0,63±0,16
Naïve Bayes	45,54±15,62	0,61±0,13	35,54±14,40	0,57±0,13
SVM: Polinomial	36,79±18,23	0,65±0,17	39,29±16,13	0,62±0,11
SVM: Polinomial norm.	43,04±11,95	0,63±0,17	45,89±18,09	0,64±0,14
SVM: Pearson VII	31,61±14,61	0,58±0,15	43,04±16,78	0,59±0,14
SVM: RBF	37,86±14,13	0,59±0,13	43,21±21,80	0,63±0,20
KNN K=1	17,68±15,76	0,54±0,08	25,36±14,48	0,51±0,11
KNN K=2	17,68±10,46	0,55±0,07	27,86±16,44	0,51±0,11
KNN K=3	10,00±11,49	0,54±0,09	25,36±11,84	0,56±0,11
KNN K=4	15,36±13,26	0,57±0,08	25,36±14,48	0,57±0,10
KNN K=5	12,68±13,19	0,57±0,08	29,29±14,95	0,60±0,10
AD: J48	37,68±18,37	0,56±0,16	40,54±19,36	0,60±0,15
AD: Random Forest N=10	33,93±14,21	0,55±0,13	48,21±18,69	0,66±0,14
AD: Random Forest N=25	39,29±11,01	0,59±0,07	41,96±20,91	0,67±0,18
AD: Random Forest N=50	35,18±16,19	0,57±0,04	44,46±22,58	0,68±0,17
AD: Random Forest N=75	35,18±15,08	0,58±0,08	43,39±24,70	0,65±0,18
AD: Random Forest N=100	38,93±18,78	0,61±0,11	39,29±19,08	0,65±0,19

4.3.2.4 **Nivel de personalización**

Los resultados de clasificar el nivel de personalización de las conversaciones se muestran en la Tabla 4.15.

Se observa que la aplicación de N-gramas ofrece mejores resultados que el uso de tokens, tanto en precisión (es mejor en 10 algoritmos frente a 8 que usan Tokens) como en el área ajo la curva ROC (13 algoritmos de los 18 funcionan mejor con N-gramas).

Las redes bayesianas con el núcleo TAN utilizando N-gramas proporcionan la mayor precisión de todas las técnicas, un 56,96 % así como un área bajo la curva ROC igual a 0,62. Similares resultados aportan las máquinas de soporte vectorial con el núcleo RBF, que obtienen el mismo valor de AUC y una precisión del 54,64 %, ligeramente inferior. Ofreciendo resultados mejores en área bajo la curva ROC, las máquinas de soporte vectorial utilizando el núcleo polinomial logran un área bajo la curva ROC de 0,65 y una precisión del 54,46 %.

La precisión menor de las obtenidas es la que proporcionan los K-vecinos más próximos usando N-gramas con K=3 y K=5, que tan solo alcanzan una precisión del 21,43 % y un AUC igual a 0,52 , y un 21,43 % y AUC igual a 0,53 respectivamente. El menor valor del área bajo la curva ROC se logra con los árboles de decisión aplicando la implementación J48, que apenas consigue un 0,46, mientras que su precisión se reduce a un 37,14 %.

4.3.2.5 **Nivel de campaña negativa**

Los resultados de la categorización del nivel de campaña negativa en las conversaciones políticas se muestra en la Tabla 4.16. En este proceso, las máquinas de soporte vectorial con el núcleo polinomial normalizado utilizando N-gramas ofrecen los mejores resultados en conjunto, con una precisión del 86,25 % y un área bajo la curva ROC de 0,6. El mismo algoritmo con palabras obtiene la misma precisión, y un valor ligeramente inferior para el AUC (0,56), por lo que se pueden considerar estadísticamente equivalentes.

En lo que respecta a precisión, las máquinas de soporte vectorial con Pearson VII logran el máximo, pero lo hacen a costa de un AUC insuficiente. En la parte inferior de la clasificación de precisión, los K-vecinos más próximos, utilizando palabras logran resultados paupérrimos, por debajo del 0,30.

Si atendemos al área bajo la curva ROC, el uso de N-gramas con SVM ofrece los mejores resultados, tal como se ha descrito anteriormente, seguido de un buen resultado logrado por las Redes Bayesianas que logran un 76,07 % de precisión y un AUC igual a 0,58, muy próximo al SVM. En cambio, KNN usando

Tabla 4.15: Resultados de la clasificación del nivel de personalización en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	49,64±13,78	0,60±0,15	49,46±12,71	0,59±0,18
RB: TAN	56,96±11,95	0,62±0,19	50,71±13,37	0,58±0,15
Naïve Bayes	51,96±12,56	0,50±0,11	36,79±20,05	0,51±0,20
SVM: Polinomial	50,71± 8,63	0,56±0,10	54,46±13,21	0,65±0,15
SVM: Polinomial norm.	51,96± 9,39	0,62±0,06	50,71±12,00	0,62±0,09
SVM: Pearson VII	42,86±16,50	0,53±0,13	48,04±13,87	0,57±0,13
SVM: RBF	49,46±12,71	0,59±0,12	54,64±15,34	0,62±0,15
KNN K=1	27,68±12,69	0,51±0,13	29,11±18,63	0,53±0,14
KNN K=2	26,61± 9,24	0,51±0,10	30,36±15,73	0,52±0,12
KNN K=3	21,43±10,14	0,52±0,13	30,18±18,64	0,51±0,17
KNN K=4	22,68±11,33	0,51±0,15	22,68±19,27	0,53±0,19
KNN K=5	21,43±10,14	0,53±0,11	26,43±13,57	0,52±0,20
AD: J48	37,14±18,80	0,46±0,18	45,36±22,65	0,59±0,19
AD: Random Forest N=10	50,36±16,10	0,56±0,12	51,96±11,09	0,53±0,16
AD: Random Forest N=25	51,96± 9,39	0,56±0,09	49,46±12,71	0,62±0,14
AD: Random Forest N=50	53,21± 9,92	0,59±0,07	45,71±11,14	0,62±0,11
AD: Random Forest N=75	50,71±10,45	0,57±0,10	48,21±11,87	0,64±0,11
AD: Random Forest N=100	51,96± 9,39	0,60±0,11	50,71±10,45	0,63±0,11

palabras con K=3 y K=4 logran los peores resultados en AUC.

En general, el uso de N-gramas y palabras conllevan resultados opuestos en términos de precisión y AUC. Es decir, N-gramas ofrece mejores resultados en precisión que palabras, y viceversa, exceptuando un par de casos.

4.3.2.6 **Marco primario**

El uso de N-gramas ofrece mejores resultados que palabras en la mayoría de los algoritmos, tanto a nivel de precisión como del área bajo la curva ROC. Respecto a la precisión, de los 18 algoritmos utilizados, 11 de ellos presentan mejores resultados cuando aplican N-gramas que si utilizan palabras; en relación al área bajo la curva ROC, 14 algoritmos mejoran si usan N-gramas respecto a si usan palabras, donde son superiores 4 de ellos.

Los mejores resultados son los ofrecidos por los árboles de decisión con bosques aletatorios y $N = 25$, logrando una precisión significativamente reducida,

Tabla 4.16: Resultados de la clasificación del nivel de campaña negativa en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	76,07±13,60	0,58±0,33	74,82±14,27	0,58±0,34
RB: TAN	77,14±17,57	0,53±0,32	77,14±15,47	0,52±0,35
Naïve Bayes	77,50±18,45	0,50±0,35	76,25±20,79	0,52±0,36
SVM: Polinomial	82,50±10,54	0,57±0,39	82,50±14,67	0,56±0,41
SVM: Polinomial norm.	86,25± 7,10	0,60±0,37	86,25± 7,10	0,56±0,37
SVM: Pearson VII	88,75± 3,95	0,36±0,39	88,75± 3,95	0,34±0,38
SVM: RBF	82,50±13,44	0,44±0,41	73,75±17,13	0,47±0,35
KNN K=1	62,32±26,82	0,41±0,22	41,96±25,41	0,37±0,29
KNN K=2	41,79±22,82	0,48±0,25	27,50±26,22	0,29±0,28
KNN K=3	38,04±24,35	0,45±0,24	21,25±17,73	0,24±0,27
KNN K=4	48,57±32,28	0,36±0,30	45,00±34,46	0,26±0,23
KNN K=5	51,07±35,38	0,41±0,34	51,25±30,87	0,37±0,28
AD: J48	78,75±14,49	0,44±0,20	77,50±16,46	0,42±0,17
AD: Random Forest N=10	81,25±10,62	0,46±0,36	85,00± 9,86	0,52±0,35
AD: Random Forest N=25	80,00±15,81	0,38±0,33	83,75±10,29	0,45±0,37
AD: Random Forest N=50	80,00±12,08	0,42±0,34	85,00± 9,86	0,50±0,38
AD: Random Forest N=75	80,00±12,08	0,42±0,34	81,07± 8,72	0,51±0,40
AD: Random Forest N=100	81,25±10,62	0,39±0,31	82,50±10,54	0,52±0,39

aunque el área bajo la curva ROC es de 0,63.

En términos de precisión, los algoritmos que mejores resultados ofrecen son los árboles de decisión con bosques aleatorios de 10 árboles (N=10) y N-gramas, que apenas llegan a un 26,61 % y si incluyen la implementación J48 cubren tan solo un 25 %. En sendos casos el área bajo la curva ROC es de 0,57 y 0.56. Con respecto al área bajo la curva ROC, los valores mayores se logran con los árboles de decisión con bosques aleatorios de 25 árboles y N-gramas, y con el núcleo basado en funciones de base radial, con los que se logra un 0,63 y una precisión de apenas el 24,11 %.

4.3.2.7 **Marco secundario**

Los resultados obtenidos de la categorización del marco secundarios son los que se muestran en la Tabla 4.18. En ellos se observa que el uso de palabras obtiene mejores valores de precisión que palabras (en 12 casos frente a 4 casos de N-

Tabla 4.17: Resultados de la clasificación del marco primario en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
RB: K2	23,93±14,83	0,62±0,17	22,68±14,06	0,63±0,16
RB: TAN	22,86±14,31	0,61±0,14	21,61±17,85	0,62±0,13
Naïve Bayes	21,25±14,49	0,56±0,10	17,68±13,37	0,56±0,08
SVM: Polinomial	15,18± 7,86	0,58±0,07	16,43±14,45	0,61±0,08
SVM: Polinomial norm.	22,86±10,02	0,63±0,06	21,61±10,50	0,62±0,08
SVM: Pearson VII	21,61± 8,69	0,58±0,14	24,29±15,59	0,60±0,10
SVM: RBF	20,36±12,29	0,58±0,11	24,11±12,52	0,63±0,06
KNN K=1	15,18± 5,21	0,52±0,04	20,18±10,41	0,54±0,08
KNN K=2	17,68±12,01	0,50±0,05	17,86± 6,99	0,50±0,06
KNN K=3	16,43±10,23	0,55±0,08	16,61± 8,91	0,49±0,07
KNN K=4	12,68± 8,35	0,50±0,06	14,11± 9,76	0,47±0,07
KNN K=5	12,86±10,74	0,50±0,06	14,11±11,40	0,50±0,09
AD: J48	21,61±10,50	0,54±0,07	25,00±14,43	0,56±0,10
AD: Random Forest N=10	26,61±14,97	0,57±0,10	23,93±12,26	0,60±0,11
AD: Random Forest N=25	24,11±11,05	0,63±0,09	20,18±13,34	0,61±0,07
AD: Random Forest N=50	24,29±16,67	0,60±0,07	18,93±13,42	0,62±0,10
AD: Random Forest N=75	19,11±10,91	0,61±0,09	21,43±16,62	0,63±0,09
AD: Random Forest N=100	17,86± 9,14	0,61±0,07	20,18±14,58	0,62±0,11

gramas y 2 indistintos). Así mismo, los N-gramas permiten alcanzar mayores resultados del área bajo la curva ROC (en 14 algoritmos, frente a 3 de palabras y en 1 caso que ofrece los mismos resultados).

El mejor resultado es el ofrecido por las redes bayesianas con el núcleo *Tree Augmented Naïve (TAN)*, que logra un 51,79 % de precisión y un área bajo la curva ROC de 0,75. Los primeros 4 algoritmos ofrecen resultados estadísticamente similares, con valores de precisión que van desde 51,96 % hasta el 47,86 %, y tienen áreas bajo la curva ROC entre 0,73 y 0,78.

En el lado opuesto, los K-vecinos más próximos utilizando N-gramas y con valores de K igual a 1, 4 y 5 ofrecen precisiones que van del 12,5 % (K=1) al 8,75 % (K=4) y del 7,5 % (K=5), con valores de AUC iguales a 0,50, 0,55 y 0,53 respectivamente.

Tabla 4.18: Resultados de la clasificación del marco secundario en conversaciones utilizando N-gramas y Tokens, en términos de precisión (%) y área bajo la curva ROC (AUC).

	N-gramas		**Tokens**	
	Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
	Precisión (%)	**AUC**	**Precisión (%)**	**AUC**
RB: K2	44,46±11,31	0,72±0,11	41,96±12,63	0,72±0,15
RB: TAN	51,79±11,87	0,75±0,13	49,46±11,26	0,73±0,12
Naïve Bayes	42,68±18,51	0,67±0,12	50,36±18,13	0,71±0,13
SVM: Polinomial	49,11±13,10	0,65±0,12	51,96±12,56	0,70±0,14
SVM: Polinomial norm.	37,86±18,40	0,68±0,14	36,61±14,71	0,73±0,10
SVM: Pearson VII	23,93±20,69	0,64±0,18	23,93±15,95	0,62±0,17
SVM: RBF	34,29±16,94	0,57±0,21	37,86±11,41	0,70±0,13
KNN K=1	12,50±19,54	0,50±0,06	50,18±20,05	0,59±0,08
KNN K=2	13,75±19,94	0,54±0,07	45,18±23,45	0,57±0,16
KNN K=3	8,75±19,59	0,55±0,07	32,68±24,28	0,58±0,12
KNN K=4	8,75±19,59	0,55±0,07	23,93±20,69	0,60±0,13
KNN K=5	7,50±19,72	0,53±0,09	20,18±22,92	0,60±0,14
AD: J48	38,21±18,37	0,59±0,14	38,21±18,37	0,57±0,14
AD: Random Forest N=10	41,61±12,80	0,60±0,14	49,82±22,50	0,68±0,16
AD: Random Forest N=25	49,11±15,53	0,67±0,14	50,89±14,37	0,72±0,14
AD: Random Forest N=50	44,29±19,65	0,67±0,17	47,86±13,59	0,78±0,09
AD: Random Forest N=75	49,29±17,82	0,67±0,14	47,86±14,82	0,78±0,10
AD: Random Forest N=100	42,86±19,40	0,67±0,13	46,79±15,40	0,78±0,11

4.3.2.8 **Resumen**

Los resultados que arroja el proceso de categorización nos permiten afirmar que es posible construir un modelo que representa algunos elementos del modelo de comportamiento comunicativo de las conversaciones políticas, aunque con determinadas limitaciones en el caso del nivel de meta-comunicación y del marco primario en dichas conversaciones.

En síntesis, tal como se expone en la Tabla 4.19, el modelo de clasificación ofrece un notable grado de calidad de los resultados en cuanto al área bajo la curva ROC. El reconocimiento de la función movilizadora, el marco secundario y la función informativa han logrado cotas elevadas en términos de calidad. En cambio, el resto de categorías no superan el 0,66. En términos de precisión, los distintos algoritmos utilizados han conseguido alcanzar resultados claramente positivos a la hora de identificar el grado de campaña negativa, la función movilizadora y la función informativa, mientras que los marcos primario y se-

cundario, además del nivel de meta-comunicación presentan niveles e dificultad mayor a la hora de obtener grados de precisión altos al ser clasificados.

Recordando la caracterización de las conversaciones a través de sus principales dimensiones -análisis funcional, el análisis argumentativo y la estrategia comunicativa-, el modelo de clasificación resulta operativa a la hora de clasificar en el ámbito del análisis funcional, mientras que padece ciertas limitaciones con respecto a las otras dos dimensiones. Esta circunstancia evidencia la complejidad de este tipo de clasificaciones, mayor en la medida en que mide términos más abstractos o basados en un contexto político determinado. En consecuencia, este modelo requiere mejoras que permitan aumentar tanto la precisión como la eficacia de la clasificación de las tres dimensiones de la acción comunicativa, a través de sus correspondientes propiedades de una conversación política.

En este punto es necesario señalar que los resultados de precisión y sensitividad obtenidos presentan considerables valores en sus desviaciones típicas, con lo que son estadísticamente menos representativos de cada clasificación. Esta circunstancia está relacionada con el reducido tamaño de los textos del cuerpo de los mensajes iniciales, así como por el efecto de acumulación de textos breves en los mensajes. Es lo que se conoce como problema de la variabilidad de los resultados datos por motivos de escasez *data sparsity problem*, que aparece a la hora de clasificar textos cortos, tal como se ha estudiado en *Twitter* [SHA12], motivo por el que se suele emplear la media [JYZ+11] como estadístico de referencia y comparación de los resultados.

De igual modo a como sucede con los mensajes políticos, no se aprecian diferencias significativas respecto a la eficacia de los tipo de tratamiento del texto -mediante N-gramas o palabras- , motivo por el cual no se puede afirmar la idoneidad de uno de ellos sobre el otro.

Tabla 4.19: Cuadro resumen con las mejores técnicas para clasificación de conversaciones en términos de precisión (%) y área bajo la curva ROC (*AUC*).

Categoría	**Técnica**	**Prec.(%)**	**AUC**
Función informativa	SVM: Polinomial normalizado (Palabras)	68,57	0,73
Función movilizadora	*SVM: Polinomial (Palabras)*	71,07	0,79
Meta-comunicación	*AD: Random Forest N=10 (Palabras)*	48,21	0,66
Personalización	*RB: TAN (N-gramas)*	56,96	0,62
Campaña negativa	SVM: Polinomial normalizado (N-gramas)	*86,25*	*0,6*
Marco primario	AD: Random Forest N=25 (N-gramas)	*24,11*	*0,63*
Marco secundario	*RB: TAN (N-gramas)*	51,79	0,75

4.4 Sumario

A lo largo de este capítulo se ha realizado la validación empírica del modelo de clasificación de mensajes y del modelo de clasificación de conversaciones. Para ello, se han repasado las técnicas a aplicar dicha validación, y se ha descrito la metodología que se ha utilizado, como han sido las técnicas de SMOTE y la validación cruzada.

Tras observar y analizar con detenimiento los resultados obtenidos a partir del modelo de representación presentado en el capítulo 3, se han presentado las conclusiones con respecto a sendos modelos de categorización de mensajes y conversaciones.

El primer modelo, relativo a la clasificación de mensajes, ha resultado funcional para su cometido, mostrando niveles elevados en términos de precisión y calidad. Este modelo se ha evidenciado como capaz de clasificar adecuadamente casi la totalidad de las variables explicativas de la acción comunicativa de los mensajes en los muros en sitios web sociales de los partidos: su función informativa, su función movilizadora, su grado de referencialidad de los contenidos, su nivel de campaña negativa, su marco secundario, su grado de contribución, su nivel de personalización y su nivel de meta-comunicación. En cambio, el modelo para clasificar el marco primario ofrece unas prestaciones inferiores.

Por su parte, la clasificación de conversaciones, validada experimentalmente en este capítulo, es capaz de categorizar en el análisis funcional con un grado de acierto notable, mientras que presenta ciertas dificultades a la hora de hacerlo en el nivel argumentativo, lo que se acentúa a nivel de estrategia comunicativa.

En síntesis, los resultados han sido relativamente positivos, pese a la dificultad de clasificar adecuadamente contenidos textuales que tienen un carácter comunicacional dependiente en gran medida del contexto político, a un nivel de abstracción alto, lo que representa una parte importante del desafío de esta investigación.

«Después de escalar una gran colina,
uno se encuentra sólo con que
hay muchas más colinas a escalar»

Nelson Mandela (1918-2013)

CAPÍTULO

5

Mejora del modelo

Como hemos visto hasta el momento, el modelo planteado en nuestra investigación, y que ha sido validado empíricamente en el capítulo 4, permite clasificar mensajes y conversaciones políticas que tienen lugar en los sitios sociales de los partidos políticos.

En particular es capaz de reconocer con cierta calidad determinados aspectos de los mensajes y de las conversaciones. En concreto, resulta eficaz para clasificar las dimensiones funcional, deliberativa, argumentativa y de estrategia comunicativa de los mensajes, aunque esta última con ciertas dificultades.

Así mismo, el modelo identifica y clasifica conversaciones políticas con relativa calidad en la dimensión funcional, algo menos la dimensión argumentativa y en menor medida la de estrategia comunicativa. No obstante, la capacidad de clasificación del nivel de meta-comunicación (en la dimensión argumentativa), el marcio primario o el marco secundario (ambos en la dimensión de estrategia comunicativa) presentan carencias notables. Por este motivo, en este capítulo se aborda una etapa de mejora que no estaba inicialmente recogida entre los objetivos específicos y operacionales descritos al inicio del documento. A lo largo de las siguientes páginas se describen las mejoras efectuadas a fin de proporcionar un modelo más completo que el diseñado y construido hasta el momento.

El capítulo se estructura tal como sigue. En la sección 5.1 presenta la técnica de *Ganancia de Información* o *Information Gain*. Esta técnica se aplica primeramente a la mejora en la categorización de los mensajes políticos publicados en los muros digitales, lo que es descrito en la sección 5.2. Seguidamente, a lo lar-

go de la sección 5.3 se emplea la misma técnica en las conversaciones políticas con dos objetivos. En primer lugar, se busca explicitar relaciones entre las clasificaciones etiquetadas para que puedean ser consideradas como características de las propias conversaciones, y de esta manera poder mejorar el proceso de clasificación. En segundo lugar, se hace uso de *Information Gain* para buscar la relevancia de cada una de las características de una conversación que se utilizan a la hora de clasificar cada una de dichas conversaciones, permitiendo afinar en el proceso de clasificación de éstas, además de la consiguiente reducción en tiempo del proceso de clasificación.

5.1 Obteniendo relevancias mediante la técnica de Ganancia de información

La técnica de *Ganancia de Información* o *Infogain* analiza la relevancia de un atributo midiendo su relación con respecto a cada clase.

Dada la dispersión de los términos textuales que aparecen tanto en el cuerpo de los mensajes políticos como en el cuerpo del *post* o mensaje inicial de las conversaciones, se ha incorporado un mecanismo que permite facilitar el proceso de entrenamiento. Este instrumento es el de la ganancia de información (procedente del término anglosajón *Information Gain* o su acrónimo *IG*), que permite conocer la relevancia de un atributo midiendo la ganancia de información respecto a la clase que lo utiliza [Ken83], a partir de esta definición:

$$IG(j) = \sum_{v_j \in R} \sum_{C_i} P(v_j, C_i) \cdot \frac{P(v_j, C_i)}{P(v_j) \cdot P(C_i)} \tag{5.1}$$

donde:

- C_i denota la clase i-ésima.
- v_j denota el valor del atributo j.
- $P(v_j, C_i)$ denota la probabilidad de que el atributo j tenga el valor v_j en la clase C_i.
- $P(v_j)$ es la probabilidad de que el atributo j tenga el valor v_j en el conjunto de datos de entrenamiento.
- $P(C_i)$ es la probabilidad de que el conjunto de datos de entrenamiento pertenezca a la clase C_i.

Mediante el empleo de esta técnica, se ha eliminado cualquier característica en el conjunto de entrenamiento con un valor de IG = 0. Como se observa a continuación en la Tabla 5.1 y en la Tabla 5.2, la aplicación del filtro IG permite eliminar los términos innecesarios para la categorización, logrando una importante reducción de ruido informacional, y cuyos efectos sobre el proceso de categorización se mostrarán en las páginas siguientes.

Los resultados muestran que se ha conseguido reducir drásticamente el número de características, facilitando así el proceso de categorización, tanto de las conversaciones como de los mensajes publicados. Será a partir de este nuevo conjunto de características ya seleccionadas sobre las que se realice una nueva etapa de categorización de ambos conjuntos de datos

Tabla 5.1: Reducción de características (%) en mensajes por aplicación de *Information Gain*.

	Núm. características		**Reducción(%)**	
	N-gramas	**Tokens**	**N-gramas**	**Tokens**
Función informativa	95	87	90,72%	91,66%
Función movilizadora	51	46	95,02%	95,59%
Metacomunicación	200	128	80,47%	87,73%
Campaña negativa	141	86	86,23%	91,75%
Nivel de Personalización	130	93	87,30%	91,08%
Grado de contribución	383	263	62,60%	74,78%
Grado de debate	201	167	80,37%	83,99%
Referencialidad de contenidos	77	83	92,48%	92,04%
Marco primario	253	191	75,29%	81,69%
Marco secundario	167	136	83,69%	86,96%

Tabla 5.2: Reducción de características (%) en conversaciones por aplicación de *Information Gain*.

	Núm. características		**Reducción(%)**	
	N-gramas	**Tokens**	**N-gramas**	**Tokens**
Función informativa	16	18	99,80%	98,47%
Función movilizadora	21	17	99,73%	98,55%
Metacomunicación	19	16	99,76%	98,64%
Campaña negativa	116	26	98,51%	97,79%
Nivel de Personalización	11	10	99,86%	99,15%
Marco primario	11	10	99,86%	99,15%
Marco secundario	22	21	99,72%	98,21%

5.2 Nuevo modelo de mensajes. Discusión de resultados

Una vez descrita la técnica de la *Ganancia de Información* o *Information Gain*, se ha aplicado en la categorización de los mensajes en los perfiles sociales de los partidos, utilizando para ello el conjunto simplificado de características, que permitirá realizar el procesamiento de manera más sencilla tratando de no perder eficacia representativa. En las líneas siguientes exponemos los resultados de dicha categorización detallada para cada una de las variables de los mensajes políticos.

5.2.1 Función informativa

El mejor clasificador de la función informativa viene del uso de *Information Gain* con árboles de decisión utilizando bosques aleatorios de 75 árboles y mediante Tokens. Esta variante nos proporciona un precisión del 97,53 % y un área bajo la curva ROC de 0,98. Si reducimos el número de árboles a 75, se consigue mantener el área, y la precisión se mantiene prácticamente igual, con un 97,52 %. Los resultados completos se muestran en la Tabla 5.3.

En un análisis comparado por tipo de término en el análisis (n-grama o palabra) con respecto al AUC, en el caso de los N-gramas, el uso de *Information Gain* da mejores resultados que si no se aplica en 11 algoritmo, igual en 6 de ellos y en uno de ellos resulta desfavorable. Otro tanto sucede al usar Tokens, en cuyo caso aplicar *Information Gain* da mejores resultados que no aplicarlo en 10 algoritmos, mientras que en 8 de ellos los resultados son idénticos. La aplicación de la *Ganancia de Información* influye en el conjunto de los algoritmos aplicados, de modo que se logra que 25 variantes, de las 72 totales, alcancen el valor máximo del área bajo la curva ROC (0,98).

En el extremo opuesto, *Naïve Bayes* obtiene los peores resultados a gran distancia del resto de algoritmos si no utiliza *Information Gain*: un AUC igual a 0,58 y una precisión del 12,61 % usando Tokens, o del 16,13 % mediante N-gramas. En cambio, si se aplica la Ganancia de Información, el mismo algoritmo llega a alcanzar un área bajo la curva ROC de 0,91 y 0,92 respectivamente.

Para el resto de algoritmos, los mayores progresos en el AUC llegan al entorno del 9 % y se producen para las redes bayesianas con el núcleo *TAN*, mientras que para el resto esta mejora no superan los 3 puntos porcentuales. Por último, en el caso de las máquinas de soporte vectorial con el núcleo *RBF*, el uso de *Information Gain* penaliza el resultado en precisión y AUC con N-gramas.

Tabla 5.3: Resultados comparados del uso de *Information Gain* para clasificar la función informativa en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	96,65±0,75	0,89±0,02	96,65±0,73	0,90±0,03
	RB: TAN	96,39±0,52	0,88±0,02	96,26±0,64	0,88±0,02
	Naïve Bayes	16,13±1,11	0,58±0,02	12,61±0,95	0,58±0,02
	SVM: Polinomial	96,45±0,69	0,97±0,02	96,45±0,64	0,98±0,01
	SVM: Polinomial norm.	96,73±0,54	0,97±0,02	96,83±0,63	0,98±0,02
	SVM: Pearson VII	96,70±0,56	0,97±0,02	96,67±0,65	0,97±0,02
	SVM: RBF	96,32±0,55	0,98±0,01	96,86±0,58	0,98±0,01
	KNN K=1	92,01±1,35	0,86±0,03	91,72±1,24	0,86±0,06
	KNN K=2	93,76±0,91	0,91±0,01	93,54±0,88	0,92±0,03
	KNN K=3	93,18±1,09	0,93±0,01	92,99±1,08	0,93±0,03
	KNN K=4	93,94±0,95	0,94±0,01	93,91±1,08	0,94±0,02
	KNN K=5	93,65±1,02	0,94±0,01	93,69±0,99	0,94±0,02
	AD: J48	97,12±0,91	0,93±0,05	97,24±0,82	0,93±0,05
	AD: Random Forest N=10	96,95±0,68	0,97±0,03	97,06±0,58	0,97±0,02
	AD: Random Forest N=25	97,11±0,62	0,97±0,02	97,08±0,53	0,98±0,01
	AD: Random Forest N=50	97,18±0,66	0,98±0,02	97,21±0,53	0,98±0,01
	AD: Random Forest N=75	97,12±0,66	0,98±0,02	97,17±0,56	0,98±0,01
	AD: Random Forest N=100	97,11±0,70	0,98±0,02	97,11±0,53	0,98±0,01
Con IG	*RB: K2*	96,76±0,71	0,95±0,03	96,83±0,76	0,95±0,03
	RB: TAN	95,97±0,47	0,94±0,02	96,00±0,52	0,95±0,02
	Naïve Bayes	52,42±1,77	0,91±0,03	48,11±3,56	0,92±0,03
	SVM: Polinomial	95,34±0,89	0,97±0,01	95,63±0,59	0,98±0,01
	SVM: Polinomial norm.	96,57±0,68	0,98±0,02	96,74±0,52	0,98±0,01
	SVM: Pearson VII	97,22±0,60	0,98±0,02	97,31±0,49	0,98±0,02
	SVM: RBF	94,99±1,01	0,97±0,02	95,33±0,68	0,98±0,01
	KNN K=1	95,16±0,71	0,89±0,03	95,31±0,98	0,89±0,04
	KNN K=2	95,60±0,70	0,94±0,02	95,60±0,70	0,94±0,02
	KNN K=3	95,24±0,69	0,95±0,02	95,27±0,84	0,95±0,02
	KNN K=4	95,53±0,62	0,96±0,02	95,47±0,84	0,96±0,02
	KNN K=5	94,76±0,57	0,96±0,02	94,81±1,02	0,96±0,02
	AD: J48	97,08±0,69	0,93±0,04	97,30±0,71	0,95±0,03
	AD: Random Forest N=10	97,22±0,68	0,97±0,01	97,34±0,74	0,97±0,02
	AD: Random Forest N=25	97,47±0,63	0,98±0,02	97,50±0,65	0,98±0,02
	AD: Random Forest N=50	97,46±0,72	0,98±0,01	97,47±0,66	0,98±0,02
	AD: Random Forest N=75	97,46±0,73	0,98±0,02	97,53±0,65	0,98±0,02
	AD: Random Forest N=100	97,41±0,75	0,98±0,02	97,52±0,65	0,98±0,02

5.2.2 **Función movilizadora**

El uso de N-gramas ofrece mejores resultados en todas los algoritmos y en todas sus variantes (con n-grama o con Tokens), tanto a nivel de precisión como en área bajo la curva ROC, tal y como se aprecia en la Table 5.4.

En términos de precisión, el mejor proviene del uso N-gramas por parte de los árboles de decisión con Random Forest y 100 árboles, que consigue una precisión del 89,06 % y un área bajo la curva ROC de 0,79. Si el número de árboles aplicados es de 75 ($N = 75$), tanto la precisión (89,01 %) como el AUC (0,78) se reducen, aunque muy ligeramente En contraste con esto, Naïve Bayes ofrece los peores datos en caso de utilizar Tokens, ya que tan solo logra una precisión del 19,12 % y un área bajo la curva ROC igual a 0,71 en caso de aplicar Information Gain o del 14,27 % y 0,58 respectivamente si no se aplica.

En lo que respecta análisis del área bajo la curva ROC, los resultados son aún mejores que los anteriores. Si se aplica Information Gain a los árboles de decisión con bosques aleatorios (ya sean 75 o 100) y se hace uso de N-gramas, se obtiene un AUC de 0,81 además de una precisión del 87,52 % y del 87,47 % respectivamente. En el otro extremo, los resultados menos favorables son los de Naïve Bayes, que sin aplicar Information Gain no logra superar el 0,58 en AUC, obteniendo precisiones del 20,79 % y del 14,27 % al utilizar N-gramas y Tokens respectivamente.

Con todo ello se observa que el uso de Information Gain mejora la maximización del área bajo la curva ROC, mientras que sin aplicarlo lo que se obtiene es una precisión mayor. En el caso de Naïve Bayes, este algoritmo que más mejora recibe por el tratamiento de Information Gain, con un incremento próximo al 50 % en precisión y casi multiplicando por dos su área bajo la curva ROC. Las principales aportaciones afectan al incremento de la curva ROC que en ningún caso quedan por debajo de 0,91 si se utilizan N-gramas o incluso del 0,94 mediante el empleo de Tokens.

La incorporación de Information Gain sobre el conjunto de datos de mensajes hace que el uso de árboles de decisión con *Random Forest* y N=75 mejore los resultados obtenidos sin Information Gain, y se convierta en el mejor clasificador, con un área bajo a curva ROC de 0,81 además de una precisión del 87,52 %. Utilizando 100 árboles en lugar de 75, alcanzamos el mismo valor del AUC, pero con una precisión un poquito menor, del 87,47 %.

Tabla 5.4: Resultados comparados del uso de *Information Gain* para clasificar la función movilizadora en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	86,50±0,72	0,67±0,04	87,11±0,51	0,70±0,04
	RB: TAN	86,19±0,79	0,64±0,03	86,34±0,90	0,65±0,03
	Naïve Bayes	20,79±2,51	0,58±0,04	14,27±1,55	0,58±0,04
	SVM: Polinomial	73,91±0,89	0,73±0,02	73,60±1,55	0,75±0,03
	SVM: Polinomial norm.	86,79±0,84	0,75±0,02	85,60±0,80	0,76±0,03
	SVM: Pearson VII	84,86±2,69	0,75±0,03	85,57±0,87	0,75±0,03
	SVM: RBF	62,05±1,89	0,76±0,03	57,63±2,01	0,77±0,03
	KNN K=1	64,03±1,57	0,62±0,03	61,39±2,48	0,62±0,02
	KNN K=2	72,39±1,47	0,64±0,02	70,20±2,10	0,65±0,03
	KNN K=3	68,77±1,91	0,65±0,03	65,99±2,58	0,67±0,03
	KNN K=4	72,68±1,66	0,65±0,03	69,99±2,13	0,68±0,03
	KNN K=5	72,10±2,18	0,65±0,03	68,81±2,53	0,68±0,03
	AD: J48	81,27±1,73	0,65±0,04	81,71±2,03	0,63±0,05
	AD: Random Forest N=10	87,55±0,96	0,73±0,05	87,61±0,51	0,74±0,04
	AD: Random Forest N=25	88,75±0,64	0,77±0,04	88,55±0,76	0,77±0,04
	AD: Random Forest N=50	88,91±0,63	0,78±0,04	88,59±0,77	0,78±0,04
	AD: Random Forest N=75	89,01±0,50	0,78±0,04	88,69±0,63	0,79±0,03
	AD: Random Forest N=100	89,06±0,51	0,79±0,03	88,66±0,41	0,79±0,04
Con IG	*RB: K2*	87,82±0,35	0,75±0,02	87,76±0,25	0,74±0,02
	RB: TAN	87,55±0,63	0,71±0,02	87,47±0,58	0,71±0,02
	Naïve Bayes	23,13±1,21	0,74±0,03	19,12±0,66	0,71±0,02
	SVM: Polinomial	56,38±1,51	0,78±0,03	53,56±2,03	0,78±0,02
	SVM: Polinomial norm.	69,69±2,84	0,78±0,03	65,64±2,95	0,78±0,03
	SVM: Pearson VII	76,13±2,01	0,77±0,03	74,96±1,55	0,77±0,03
	SVM: RBF	49,36±1,52	0,76±0,02	48,58±1,66	0,76±0,02
	KNN K=1	75,47±1,41	0,66±0,03	74,65±1,27	0,66±0,03
	KNN K=2	80,53±1,00	0,70±0,03	79,96±1,30	0,71±0,04
	KNN K=3	74,39±1,27	0,73±0,03	74,01±1,40	0,73±0,03
	KNN K=4	76,90±1,22	0,73±0,03	75,95±1,16	0,74±0,03
	KNN K=5	73,37±1,26	0,74±0,03	72,90±1,54	0,75±0,03
	AD: J48	82,07±1,02	0,66±0,03	82,80±1,43	0,67±0,02
	AD: Random Forest N=10	86,66±0,73	0,78±0,02	86,27±0,67	0,77±0,03
	AD: Random Forest N=25	86,92±1,09	0,80±0,03	86,52±1,04	0,79±0,03
	AD: Random Forest N=50	87,30±0,87	0,81±0,03	86,82±0,81	0,80±0,03
	AD: Random Forest N=75	87,52±0,96	0,81±0,03	87,07±0,86	0,80±0,03
	AD: Random Forest N=100	87,47±1,00	0,81±0,03	86,87±0,84	0,80±0,03

5.2.3 Nivel de meta-comunicación

Como se observa en la Tabla 5.5, la aplicación de la técnica de *Ganancia de Información* es desigual en cuanto a mejoras aportadas. Cuando se trabaja con N-gramas, aporta incrementos de precisión en la mitad de los casos, mientras que cuando se utilizan Tokens sí mejora (obtiene mejores valores en 12 de los 18 casos, y en 4 el resultado no varía). En términos de área bajo la curva ROC, incorporar *Information Gain* cuando se emplean N-gramas aumenta la precisión en 16 de los 18 casos (en los otros dos queda igual). Por otro lado, si se emplean Tokens, 12 algoritmos mejoran, mientras que 4 se mantienen y 2 obtienen incluso peor resultado.

En relación con la precisión, cuando las familias *KNN* y *SVM* utilizan N-gramas, salen desfavorecidas por la utilización de *Information Gain*, en tanto en cuanto obtienen peores resultados que si no se aplicase. En cambio, si utilizan Tokens, las máquinas de soporte vectorial o *SVM* obtienen peores resultados si se aplica *Ganancia de Información*. En términos de AUC, casi todos los algoritmos obtienen mejoras, siendo el más beneficiado *Naïve Bayes* proporcionalmente (que es el más beneficiado en general, partiendo de los peores resultados sin *Information Gain*). Por contra, las máquinas de soporte vectorial con *Pearson VII* y *RBF* retroceden en AUC cuando emplean Tokens.

El mejor resultado global es el de los árboles de decisión con *Random Forest* de 100 árboles utilizando N-gramas que tiene una precisión del 79,18 % y un valor del área bajo la curva ROC de 0,81. En el mismo entorno se encuentran los K-vecinos más próximos con K=1 (tanto usando N-gramas como Tokens), y los árboles de decisión con bosques aleatorios con 50, 75 y 100 árboles que emplean N-gramas, o los mismos tipos de árboles con N=25 y N=75 que emplean Tokens.

En cuanto al área bajo la curva ROC, el uso de *Ganancia de Información* y N-gramas permite a los árboles de decisión con bosques aleatorios de 100 y 50 árboles obtener el valor máximo, igual a 0,81, con precisiones del 79,18 % y 78,95 % respectivamente.

Al contrario, *Naïve Bayes* sin aplicar *Ganancia de Información* presenta los peores resultados, tanto en precisión (un 16,1 % con N-gramas y un 10,3 % con Tokens) como en área bajo la curva ROC (0,53 empleando N-gramas y 0,52 mediante Tokens).

Tabla 5.5: Resultados comparados del uso de *Information Gain* para clasificar el nivel de meta-comunicación en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	75,56±0,72	0,69±0,02	75,72±0,82	0,70±0,02
	RB: TAN	74,73±1,01	0,67±0,02	74,68±1,08	0,68±0,02
	Naïve Bayes	16,10±1,29	0,53±0,02	10,30±1,04	0,52±0,02
	SVM: Polinomial	66,49±1,40	0,77±0,01	65,70±1,39	0,77±0,01
	SVM: Polinomial norm.	77,79±1,23	0,77±0,03	77,63±1,27	0,77±0,02
	SVM: Pearson VII	78,16±1,08	0,79±0,02	77,09±1,14	0,79±0,02
	SVM: RBF	58,70±1,56	0,79±0,02	56,32±1,79	0,79±0,01
	KNN K=1	62,21±1,31	0,64±0,01	58,86±2,12	0,65±0,02
	KNN K=2	68,74±0,97	0,66±0,01	65,93±1,71	0,67±0,02
	KNN K=3	67,20±1,18	0,67±0,02	63,59±1,53	0,68±0,02
	KNN K=4	69,26±0,96	0,68±0,02	65,98±1,48	0,69±0,02
	KNN K=5	69,26±1,17	0,68±0,02	65,90±1,79	0,69±0,02
	AD: J48	70,55±1,28	0,66±0,03	70,94±1,16	0,67±0,02
	AD: Random Forest N=10	76,48±1,34	0,76±0,02	76,58±1,15	0,76±0,02
	AD: Random Forest N=25	77,75±1,07	0,78±0,02	78,31±1,09	0,79±0,02
	AD: Random Forest N=50	78,45±0,81	0,79±0,02	78,54±1,01	0,80±0,02
	AD: Random Forest N=75	78,79±0,75	0,80±0,02	78,93±1,11	0,80±0,02
	AD: Random Forest N=100	78,73±0,73	0,80±0,02	78,96±1,10	0,80±0,02
Con IG	*RB: K2*	76,79±0,74	0,73±0,02	76,52±0,83	0,74±0,02
	RB: TAN	77,31±0,72	0,69±0,01	77,22±1,08	0,71±0,01
	Naïve Bayes	24,24±2,47	0,66±0,02	24,73±1,80	0,63±0,02
	SVM: Polinomial	59,68±1,40	0,80±0,02	57,88±1,58	0,79±0,01
	SVM: Polinomial norm.	68,91±1,72	0,78±0,02	65,52±1,10	0,78±0,02
	SVM: Pearson VII	73,85±0,66	0,79±0,01	71,95±0,95	0,78±0,02
	SVM: RBF	49,23±1,49	0,79±0,01	48,94±1,75	0,77±0,01
	KNN K=1	62,45±1,13	0,66±0,01	63,27±1,08	0,66±0,02
	KNN K=2	67,22±0,95	0,70±0,01	68,21±1,14	0,71±0,02
	KNN K=3	62,16±0,73	0,72±0,01	63,29±1,34	0,73±0,02
	KNN K=4	62,73±1,15	0,73±0,01	64,27±1,52	0,74±0,02
	KNN K=5	60,94±1,24	0,74±0,01	61,93±0,85	0,75±0,02
	AD: J48	70,05±1,00	0,67±0,02	70,26±1,44	0,67±0,02
	AD: Random Forest N=10	76,74±1,24	0,78±0,02	76,42±1,10	0,77±0,01
	AD: Random Forest N=25	78,04±0,59	0,80±0,02	77,95±1,10	0,79±0,02
	AD: Random Forest N=50	78,95±0,82	0,81±0,02	78,22±1,11	0,80±0,02
	AD: Random Forest N=75	78,90±1,11	0,81±0,02	78,68±1,14	0,80±0,02
	AD: Random Forest N=100	79,18±1,14	0,81±0,02	78,50±1,30	0,81±0,02

5.2.4 Nivel de personalización

La Tabla 5.6 muestra los resultados de la categorización del nivel de personalización en los mensajes. En ella se observa que la aplicación de Information Gain mejora la precisión en cuanto a precisión obtenida. Cuando se emplean N-gramas, 10 algoritmos obtienen mejores resultados, mientras que 8 de ellos retroceden. De manera similar, cuando se trata de Tokens, 11 algoritmos alcanzan mayor precisión, mientras que 7 la pierden.

En lo que respecta al área bajo la curva ROC, la incorporación de *Information Gain* provoca una mejora comparativa, de tal modo que al utilizan N-gramas, 16 algoritmos logran mejores resultados frente a 1 que obtiene idéntico valor y otro que pierde AUC. Otro tanto sucede cuando lo que se utilizan son Tokens, cuando 16 algoritmos avanzan, mientras que 2 obtienen peores valores si se aplica *Information Gain*.

Por familias de algoritmos, los más beneficiados por el uso de *Information Gain* son los *K*-vecinos más próximos en cualquiera de las modalidades definidas en la experimentación. A los que peor afecta la aplicación de esta técnica es a las máquinas de soporte vectorial con el núcleo *RBF*, que obtienen peores áreas bajo la curva ROC tanto con N-gramas como con Tokens. En cuanto a precisión, o se quedan igual o retroceden, pero en ningún caso avanzan. Algo similar sucede a *Naïve Bayes* que, aunque obtiene resultados positivos en AUC, en cambio pierde precisión.

Los mejores valores de precisión y área bajo la curva ROC se obtienen aplicando *Information Gain* y utilizando Tokens a los árboles de decisión con *Random Forest* y 75 ó 100 árboles. En estos dos casos se alcanzan unas precisiones del 80,8 % y 80,8 % respectivamente. En sendos casos, el valor del área sobre la curva ROC llega al 0,88 y 0,89. Los *K*-vecinos más próximos con $K=1$ sin aplicar *Ganancia de Información* presentan los resultados más pobres del conjunto, que aun así, superan el 60 % de precisión (62,51 % con Tokens y 61,77 % con N-gramas) con valores del área bajo la curva ROC de 0,59 con Tokens y 0,6 con N-gramas. Como se aprecia, la incorporación de la técnica de *Information Gain* aporta mejoras a los resultados, y sobre todo afecta muy favorablemente a los algoritmos de *K*-vecinos más próximos y *Naïve Bayes*.

5.2.5 Grado de contribución

En la Tabla 5.7 aparecen los resultados comparativos de la categorización del grado de contribución. En esta tabla se observa el desigual efecto que la aplicación de *Information Gain* ha tenido sobre los distintos algoritmos. En relación

Tabla 5.6: Resultados comparados del uso de *Information Gain* para clasificar el nivel de personalización en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	65,89±1,14	0,78±0,02	65,49±2,20	0,78±0,02
	RB: TAN	67,67±1,82	0,80±0,02	67,44±2,50	0,81±0,02
	Naïve Bayes	26,25±2,50	0,65±0,02	28,47±1,69	0,64±0,02
	SVM: Polinomial	66,60±1,90	0,81±0,01	68,15±1,95	0,81±0,02
	SVM: Polinomial norm.	74,40±1,47	0,84±0,01	76,36±1,71	0,86±0,01
	SVM: Pearson VII	72,52±2,00	0,81±0,01	75,25±2,15	0,83±0,01
	SVM: RBF	51,57±3,83	0,77±0,02	49,99±4,52	0,78±0,02
	KNN K=1	61,77±2,48	0,60±0,01	62,51±2,50	0,59±0,02
	KNN K=2	61,32±2,39	0,61±0,02	61,64±2,20	0,61±0,02
	KNN K=3	59,55±3,10	0,61±0,01	60,63±2,91	0,61±0,02
	KNN K=4	60,44±3,24	0,62±0,01	62,45±3,05	0,62±0,02
	KNN K=5	59,71±3,38	0,62±0,01	62,54±3,25	0,62±0,02
	AD: J48	72,05±2,11	0,77±0,02	72,56±1,76	0,78±0,01
	AD: Random Forest N=10	72,39±2,51	0,81±0,02	71,92±2,21	0,81±0,02
	AD: Random Forest N=25	76,35±1,74	0,85±0,01	76,36±1,31	0,85±0,01
	AD: Random Forest N=50	77,85±1,65	0,86±0,01	78,17±0,89	0,86±0,01
	AD: Random Forest N=75	78,85±1,69	0,87±0,01	78,85±1,10	0,87±0,01
	AD: Random Forest N=100	78,96±1,66	0,87±0,01	79,33±1,00	0,87±0,01
Con IG	*RB: K2*	63,89±1,96	0,80±0,01	63,26±2,41	0,80±0,02
	RB: TAN	66,87±1,87	0,83±0,01	67,49±2,02	0,83±0,01
	Naïve Bayes	25,65±1,38	0,74±0,03	24,44±1,17	0,75±0,03
	SVM: Polinomial	70,56±1,88	0,86±0,02	72,59±2,61	0,86±0,01
	SVM: Polinomial norm.	74,23±1,41	0,87±0,01	75,47±2,00	0,88±0,01
	SVM: Pearson VII	74,70±1,33	0,87±0,01	76,74±1,97	0,87±0,01
	SVM: RBF	39,18±8,20	0,77±0,02	37,59±8,62	0,77±0,02
	KNN K=1	66,70±1,35	0,71±0,01	68,24±1,45	0,72±0,01
	KNN K=2	64,37±1,38	0,73±0,01	65,17±1,73	0,75±0,01
	KNN K=3	60,77±1,80	0,75±0,01	61,87±1,70	0,76±0,01
	KNN K=4	59,78±2,09	0,75±0,01	60,89±1,82	0,77±0,01
	KNN K=5	57,94±1,77	0,75±0,01	58,83±2,09	0,77±0,01
	AD: J48	69,79±1,69	0,76±0,01	70,69±2,63	0,77±0,02
	AD: Random Forest N=10	76,49±1,94	0,85±0,01	77,30±1,50	0,85±0,01
	AD: Random Forest N=25	78,79±1,94	0,87±0,01	79,63±1,19	0,87±0,01
	AD: Random Forest N=50	80,13±2,04	0,88±0,01	80,38±1,26	0,88±0,01
	AD: Random Forest N=75	80,16±1,71	0,88±0,01	80,80±1,33	0,88±0,01
	AD: Random Forest N=100	80,44±1,52	0,88±0,01	80,86±1,37	0,89±0,01

con la precisión, dos terceras partes han empeorado sus resultados, ya sean empleando N-gramas o Tokens, mientras que una tercera parte ha incrementado su precisión. En términos de área bajo la curva ROC, estos resultados son más heterogéneos. Cuando se emplean N-gramas, 10 algoritmos mejoran, mientras que 4 quedan igual y otros 4 obtienen peor valor. Si se trata de Tokens, 8 algoritmos avanzan, otros 8 se mantienen y 2 retroceden.

Los algoritmos que más han mejorado su resultado, tanto en precisión como en área bajo la curva ROC han sido los *K*-vecinos más próximos, al igual que redes bayesianas *Naïve Bayes*, partiendo ambos grupos de los resultados más flojos. En cambio, aquellos que partían de buenas números han retorcedido o se han estancado. En cuanto a los efectos producidos, los progresos han afectado más al área bajo la curva ROC que a la precisión, que aun así se ha visto incrementada.

El uso de Tokens ofrece resultados algo mejores que el uso de N-gramas en términos de precisión, mientras que en el área bajo la curva ROC, sucede al revés: 9 algoritmos con N-gramas obtienen más AUC que sus equivalentes usando Tokens; 3 son mejores si usan Tokens que si emplean N-gramas y, por último, 24 de ellos ofrecen el mismo AUC usando Tokens que al trabajar con N-gramas.

La técnica que mayores valores de precisión aporta en relación con un área bajo la curva ROC elevado son las redes bayesianas con el núcleo *TAN*, que logran un 82,34% de precisión y un AUC igual a 0,81, tanto en las modalidades de N-gramas como de Tokens, y en ambos casos sin la aplicación de *Ganancia de Información*. En este entorno de resultados hay un nutrido grupo de variantes que alcanzan valores de precisión y área bajo la curva ROC bastante similares, esto es, que estadísticamente no presentan variaciones significativas. Con valores algo similares, los máximos en AUC se obtienen mediante árboles de decisión con *Random Forest* con 100 y 75 árboles, llegando a un 0,83 ambos, y unas precisiones del 82.32% y 82,24% respectivamente. En ambos casos se han empleado N-gramas.

5.2.6 **Referencialidad de contenidos**

Los resultados comparativos de aplicar *Ganancia de Información* a se muestran en la Tabla 5.8. En ella se observa que su efecto en relación con la precisión supone un retroceso parcial. Cuando se utilizan N-gramas, en 11 de los casos se obtiene peor resultado con *Information Gain* que sin él (en los otros 7 se producen mejoras), mientras que cuando se emplean Tokens, se obtienen peores resultados en 12 de los casos, en tanto que se mejora en 6 de ellos. En contraste

Tabla 5.7: Resultados comparados del uso de *Information Gain* para clasificar el grado de contribución en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	78,48±2,23	0,80±0,03	79,84±1,98	0,80±0,03
	RB: TAN	82,34±1,76	0,81±0,02	82,34±1,30	0,81±0,02
	Naïve Bayes	51,38±2,25	0,73±0,03	43,84±1,79	0,72±0,03
	SVM: Polinomial	73,60±2,10	0,77±0,03	73,92±2,59	0,77±0,02
	SVM: Polinomial norm.	80,18±1,96	0,78±0,03	79,84±1,88	0,78±0,03
	SVM: Pearson VII	81,20±1,79	0,78±0,03	81,23±1,61	0,78±0,03
	SVM: RBF	72,33±2,58	0,78±0,03	69,23±2,54	0,77±0,03
	KNN K=1	47,89±1,79	0,58±0,02	47,39±1,83	0,57±0,01
	KNN K=2	37,56±1,32	0,60±0,02	36,70±1,47	0,60±0,02
	KNN K=3	45,03±1,92	0,62±0,02	44,81±1,68	0,62±0,02
	KNN K=4	38,01±1,29	0,63±0,02	37,12±2,08	0,63±0,02
	KNN K=5	43,71±1,69	0,64±0,02	42,46±1,67	0,65±0,02
	AD: J48	76,51±1,77	0,72±0,03	76,16±2,05	0,72±0,03
	AD: Random Forest N=10	78,99±2,39	0,79±0,03	79,56±2,12	0,79±0,03
	AD: Random Forest N=25	81,74±2,11	0,81±0,02	81,90±1,89	0,81±0,02
	AD: Random Forest N=50	81,89±2,07	0,82±0,03	82,12±2,14	0,82±0,02
	AD: Random Forest N=75	82,24±1,85	0,83±0,03	82,28±2,06	0,82±0,02
	AD: Random Forest N=100	82,32±1,73	0,83±0,03	82,24±2,15	0,82±0,02
Con IG	*RB: K2*	68,04±2,16	0,80±0,03	69,00±2,64	0,80±0,03
	RB: TAN	73,59±1,93	0,81±0,02	74,65±2,60	0,81±0,02
	Naïve Bayes	62,00±3,32	0,75±0,03	60,42±1,73	0,74±0,03
	SVM: Polinomial	70,97±2,45	0,79±0,03	70,66±2,55	0,79±0,03
	SVM: Polinomial norm.	75,73±1,95	0,79±0,03	73,63±1,91	0,79±0,02
	SVM: Pearson VII	80,12±1,90	0,80±0,02	78,60±1,48	0,78±0,02
	SVM: RBF	65,51±2,62	0,77±0,03	62,37±2,40	0,74±0,03
	KNN K=1	63,05±1,94	0,65±0,03	65,23±1,00	0,65±0,01
	KNN K=2	54,26±2,98	0,67±0,04	57,72±1,14	0,67±0,01
	KNN K=3	52,35±2,54	0,68±0,03	56,49±1,89	0,69±0,01
	KNN K=4	49,03±2,15	0,69±0,03	52,83±1,51	0,70±0,01
	KNN K=5	50,05±2,09	0,70±0,03	54,46±1,43	0,70±0,01
	AD: J48	73,15±1,52	0,71±0,02	72,64±1,98	0,71±0,03
	AD: Random Forest N=10	74,83±1,67	0,80±0,02	75,22±2,36	0,79±0,03
	AD: Random Forest N=25	78,85±2,09	0,81±0,02	79,45±2,29	0,81±0,03
	AD: Random Forest N=50	79,11±2,29	0,82±0,03	79,96±2,09	0,82±0,03
	AD: Random Forest N=75	79,94±2,53	0,82±0,03	80,53±2,15	0,82±0,03
	AD: Random Forest N=100	79,91±2,21	0,82±0,03	80,42±1,92	0,82±0,03

con lo anterior, la aplicación de *Information Gain* sí implica mejores resultados en el área bajo la curva ROC, tanto en N-gramas como en Tokens en todos los algoritmos, con la salvedad de tan solo un caso en el que el AUC se mantiene.

El mayor avance lo ha experimentado las redes bayesianas Naïve Bayes. En precisión, han salido ganando los árboles de decisión, tanto en N-gramas como con Tokens, mientras que obtienen resultados ligeramente inferiores en área bajo la curva ROC. En cambio, los mayores retrocesos son para las máquinas de soporte vectorial que obtienen resultados significativamente peores en precisión, no así en AUC.

Los mejores resultados en precisión son los obtenidos por los árboles de decisión. El primer lugar lo ocupan los árboles de decisión con bosques aleatorios de 100 árboles aplicando *Information Gain* y n-grams, que logra una precisión del 92,53 % y un área bajo la curva ROC de 0,9. En un entorno estadísticamente similar se encuentran 23 de las posibles variantes de algoritmos experimentadas, principalmente los árboles de decisión. Los primeros lugares los ocupan aquellos casos que aplican *Information Gain*, a los que señaden otros sin *IG*, además de las máquinas de soporte vectorial con Pearson VII que no aplican *Ganancia de Información*.

5.2.7 **Grado de debate político**

Los resultados comparados de aplicación de *Ganancia de Información* a la clasificación del grado de debate político se muestran en la Tabla 5.9.

Cuando se aplica esta técnica, 14 algoritmos mejoran su precisión, mientras que 4 obtienen resultados inferiores. Así mismo, el área bajo la curva ROC aumenta por uso de IG en 15 algoritmos, en 2 los resultados son iguales que sin emplear IG y en uno de ellos se obtiene un AUC inferior. Estos resultados en precisión y área bajo la curva ROC son idénticos cuando se trata de N-gramas que con Tokens.

La familia de algoritmos *K*-vecinos más cercanos es la que proporcionalmente mayor aumento experimenta, especialmente en precisión. En cambio, las máquinas de soporte vectorial obtiene valores de precisión inferior cuando emplea N-gramas y en casi todos los casos en que usa Tokens. También experimentan retroceso en el área bajo la curva ROC cuando el núcleo es polinomial normalizado.

Los mejores resultados, de manera individual, son los alcanzados por los árboles de decisión con Random Forest y 100 árboles, empleando N-gramas y aplicando *Ganancia de Información*. Esta técnica obtiene una precisión del 65,13 %

Tabla 5.8: Resultados comparados del uso de *Information Gain* para clasificar la referencialidad de los contenidos en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	83,62±2,44	0,81±0,04	84,16±1,76	0,82±0,04
	RB: TAN	87,36±1,41	0,81±0,04	87,31±2,12	0,83±0,05
	Naïve Bayes	33,78±2,88	0,73±0,03	34,60±2,96	0,75±0,02
	SVM: Polinomial	75,38±1,64	0,80±0,01	75,66±2,03	0,80±0,03
	SVM: Polinomial norm.	87,67±0,98	0,82±0,02	87,33±1,34	0,82±0,02
	SVM: Pearson VII	89,48±1,21	0,78±0,06	89,91±0,94	0,82±0,03
	SVM: RBF	55,11±1,82	0,84±0,02	52,01±2,04	0,84±0,02
	KNN K=1	81,58±1,24	0,76±0,02	80,51±1,94	0,75±0,03
	KNN K=2	74,89±2,73	0,78±0,02	72,97±3,28	0,76±0,03
	KNN K=3	68,93±3,34	0,78±0,02	65,93±3,64	0,77±0,03
	KNN K=4	63,75±3,34	0,78±0,02	60,10±4,36	0,77±0,03
	KNN K=5	59,53±4,06	0,79±0,03	54,68±5,35	0,78±0,03
	AD: J48	83,26±1,18	0,81±0,02	85,57±1,37	0,83±0,02
	AD: Random Forest N=10	89,25±0,60	0,84±0,02	88,93±0,87	0,84±0,02
	AD: Random Forest N=25	91,57±0,81	0,86±0,02	91,26±0,73	0,86±0,02
	AD: Random Forest N=50	91,80±0,69	0,87±0,02	91,98±0,52	0,87±0,02
	AD: Random Forest N=75	91,94±0,76	0,87±0,02	91,99±0,70	0,87±0,02
	AD: Random Forest N=100	91,89±0,70	0,87±0,02	91,98±0,63	0,87±0,02
Con IG	*RB: K2*	84,85± 1,47	0,88±0,03	83,45± 1,46	0,88±0,03
	RB: TAN	86,03± 2,27	0,87±0,03	86,63± 1,86	0,87±0,05
	Naïve Bayes	78,95±10,19	0,82±0,02	55,85±11,77	0,82±0,02
	SVM: Polinomial	59,52± 3,12	0,86±0,02	59,14± 2,33	0,86±0,02
	SVM: Polinomial norm.	68,08± 1,93	0,86±0,02	68,69± 1,40	0,86±0,02
	SVM: Pearson VII	79,91± 2,22	0,86±0,03	79,97± 2,44	0,85±0,03
	SVM: RBF	50,37± 3,85	0,85±0,02	46,85± 2,95	0,84±0,03
	KNN K=1	80,69± 1,22	0,76±0,02	80,35± 1,61	0,76±0,02
	KNN K=2	72,21± 1,47	0,78±0,02	72,46± 1,56	0,78±0,02
	KNN K=3	64,85± 1,04	0,79±0,02	65,03± 1,46	0,79±0,03
	KNN K=4	58,77± 1,55	0,80±0,02	58,87± 1,57	0,80±0,03
	KNN K=5	54,20± 1,74	0,81±0,02	53,95± 1,51	0,80±0,03
	AD: J48	83,01± 1,64	0,84±0,02	83,62± 1,82	0,84±0,03
	AD: Random Forest N=10	90,07± 1,03	0,88±0,02	89,99± 1,12	0,88±0,02
	AD: Random Forest N=25	92,20± 0,79	0,89±0,03	91,83± 0,66	0,89±0,02
	AD: Random Forest N=50	92,49± 0,58	0,89±0,02	92,29± 0,81	0,89±0,02
	AD: Random Forest N=75	92,48± 0,49	0,89±0,02	92,46± 0,83	0,89±0,02
	AD: Random Forest N=100	92,53± 0,61	0,90±0,02	92,45± 0,72	0,89±0,02

y un área bajo la curva ROC de 0,78. Un grupo formado por 18 algoritmos forman el grupo de cabeza en cuanto a precisión y área bajo la curva ROC. En su mayoría incluyen la aplicación de *Information Gain*, y son árboles de decisión

con bosques aleatorios y un número elevado de árboles. También forman parte de este grupo las redes bayesianas con núcleos *TAN* y *K2* empleando ambas *IG*.

Por su parte, y pese a las mejoras alcanzadas por incorporación de *Information Gain*, los algoritmos *K*-vecinos más próximos se sitúan en la parte baja de la clasificación por resultados de precisión y de AUC. Finalmente, Naïve Bayes presenta los resultados más pobres en el global, tanto sin aplicar *Ganancia de Información* como aplicándola.

5.2.8 **Nivel de campaña negativa**

Los resultados de la nueva clasificación del nivel de campaña política aplicando *Information Gain* se muestran en la Tabla 5.10. La aplicación de esta técnica mejora la precisión de los alcanzados sin ella con ciertos matices. En el caso de emplearse N-gramas, 10 algoritmos obtienen mejores resultados, mientras que 8 de ellos consiguen menos precisión con *Information Gain*. En cambio, cuando se usan Tokens, son 13 los que mejoran frente a 5 que ofrecen peores valores de precisión. Por el contrario, la aplicación de la *Ganancia de Información* mejora el área bajo la curva ROC en todos los casos excepto en uno de ellos al emplear Tokens, en el cual obtiene un resultado menor.

Por algoritmos, el mayor avance porcentual es el logrado por *Naïve Bayes*, tanto en precisión como en área bajo la curva ROC, y ya sea empleando *n*-gramas o Tokens. El resto de redes bayesianas (los correspondientes a los núcleos *K2* y *TAN*) experimentan avances en el área bajo la curva ROC estadísticamente significativos, al tiempo que la mejoría en precisión es poco relevante. En AUC el grupo que ofrece una mejora porcentual es el de *K*-vecinos más próximos; esta mejora no se traslada a la precisión, donde se encontraban en la zona media antes de aplicar IG.

En cuanto a resultados, el mejor de todos es el alcanzado después de introducir *Information Gain* a los árboles de decisión con bosques aleatorios con 100 árboles y empleando Tokens. En este caso, la precisión alcanzada llega al 90,84 % y describiendo un área bajo la curva ROC con valor igual a 0,89. Esta primera posición es compartida por un grupo notable de los algoritmos -22 en total-, que se encuentran en la misma zona de significación del resultado mejor en decisión y área bajo la curva ROC. Este conjunto de algoritmos incluye a los árboles de decisión con *Random Forest* de más de 10 árboles, así como las máquinas de soporte vectorial con núcleo polinomial normalizado y con *Pearson VII* sin aplicarles *Information Gain*.

En el grupo de cola, *Naïve Bayes* muestra los resultados más débiles, ya sea

Tabla 5.9: Resultados comparados del uso de *Information Gain* para clasificar el grado de debate político en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	45,83±2,00	0,71±0,01	47,30±1,30	0,72±0,01
	RB: TAN	54,58±1,76	0,75±0,01	54,24±1,56	0,76±0,02
	Naïve Bayes	18,66±1,08	0,63±0,01	23,10±1,81	0,64±0,01
	SVM: Polinomial	48,52±1,79	0,73±0,01	49,28±1,98	0,74±0,02
	SVM: Polinomial norm.	58,47±2,17	0,76±0,01	59,11±1,42	0,77±0,01
	SVM: Pearson VII	59,40±5,21	0,74±0,02	58,83±2,40	0,75±0,01
	SVM: RBF	33,79±2,67	0,73±0,02	31,43±2,97	0,73±0,02
	KNN K=1	39,88±2,01	0,59±0,02	40,10±1,97	0,60±0,02
	KNN K=2	31,53±1,90	0,61±0,01	31,29±2,56	0,61±0,02
	KNN K=3	30,08±1,90	0,62±0,01	29,13±2,43	0,62±0,02
	KNN K=4	26,50±1,63	0,63±0,01	25,64±1,71	0,62±0,02
	KNN K=5	25,68±1,89	0,63±0,01	24,24±1,46	0,62±0,01
	AD: J48	48,69±2,14	0,65±0,01	49,70±1,72	0,66±0,02
	AD: Random Forest N=10	52,78±1,33	0,72±0,02	52,91±2,37	0,72±0,02
	AD: Random Forest N=25	57,27±1,68	0,75±0,01	57,34±1,77	0,75±0,02
	AD: Random Forest N=50	59,55±1,67	0,77±0,01	59,90±1,35	0,77±0,02
	AD: Random Forest N=75	60,16±1,53	0,77±0,01	60,18±1,44	0,77±0,02
	AD: Random Forest N=100	60,56±1,34	0,78±0,01	60,67±1,62	0,78±0,02
Con IG	*RB: K2*	62,99±1,10	0,72±0,02	63,32±0,87	0,73±0,02
	RB: TAN	65,10±0,59	0,75±0,01	65,07±0,57	0,76±0,02
	Naïve Bayes	18,98±1,80	0,64±0,02	20,92±1,38	0,64±0,02
	SVM: Polinomial	46,05±1,52	0,76±0,01	46,11±2,18	0,77±0,01
	SVM: Polinomial norm.	51,67±1,45	0,75±0,02	50,62±1,39	0,76±0,01
	SVM: Pearson VII	58,12±1,31	0,76±0,01	58,19±1,55	0,77±0,01
	SVM: RBF	32,61±1,86	0,74±0,02	32,78±1,72	0,74±0,02
	KNN K=1	46,15±2,21	0,63±0,02	47,95±2,80	0,64±0,02
	KNN K=2	44,59±2,52	0,66±0,01	46,74±2,21	0,67±0,02
	KNN K=3	44,43±1,85	0,67±0,01	46,09±2,30	0,69±0,02
	KNN K=4	43,01±1,70	0,68±0,01	45,10±2,34	0,69±0,02
	KNN K=5	43,29±1,35	0,69±0,01	45,58±1,72	0,70±0,02
	AD: J48	54,78±1,27	0,66±0,02	55,91±1,35	0,67±0,01
	AD: Random Forest N=10	60,18±1,25	0,73±0,01	61,78±1,70	0,75±0,02
	AD: Random Forest N=25	63,26±1,18	0,77±0,01	64,24±0,95	0,77±0,02
	AD: Random Forest N=50	64,79±0,89	0,78±0,01	64,75±0,95	0,78±0,02
	AD: Random Forest N=75	65,11±1,02	0,78±0,01	64,78±1,02	0,79±0,02
	AD: Random Forest N=100	65,13±1,19	0,78±0,01	64,62±0,70	0,79±0,01

empleando *Ganancia de Información* o no, y con N-gramas o Tokens, con valores de precisión que van desde el 28,9 % y un valor de AUC de 0,71 cuando se aplica IG usando Tokens hasta el mínimo, con una precisión de apenas el 10,09 % y

un AUC de 0,52, cuando no se aplica *Ganancia de Información* y se emplean Tokens. Entre ambas, las dos opciones intermedias usando N-gramas consiguen un 27,8 % de precisión y un AUC igual a 0,76 cuando incluye *Information Gain*, o del 18,12 % y un área bajo la curva ROC de 0,55 cuando no lo incluye.

Tabla 5.10: Resultados comparados del uso de *Information Gain* para clasificar el nivel de campaña negativa en mensajes, en términos de precisión(%) y AUC.

		N-gramas		**Tokens**	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	85.77±0.55	0.66±0.03	86.18±0.58	0.72±0.04
	RB: TAN	86.85±0.40	0.65±0.03	86.49±0.72	0.67±0.03
	Naïve Bayes	18.12±0.75	0.55±0.04	10.09±0.74	0.52±0.04
	SVM: Polinomial	82.22±1.79	0.82±0.03	81.74±1.30	0.82±0.02
	SVM: Polinomial norm.	89.55±0.88	0.84±0.03	89.10±0.81	0.85±0.02
	SVM: Pearson VII	89.67±0.85	0.84±0.02	89.44±0.64	0.84±0.02
	SVM: RBF	82.69±1.52	0.85±0.02	81.49±1.41	0.85±0.02
	KNN K=1	76.26±1.57	0.63±0.02	74.22±1.35	0.64±0.01
	KNN K=2	82.45±0.86	0.64±0.03	80.23±1.30	0.65±0.02
	KNN K=3	80.38±0.86	0.64±0.03	77.93±1.71	0.65±0.02
	KNN K=4	82.72±1.30	0.64±0.03	80.31±1.59	0.66±0.02
	KNN K=5	82.13±1.10	0.64±0.03	80.04±1.54	0.66±0.01
	AD: J48	87.22±0.59	0.72±0.03	88.11±1.15	0.76±0.03
	AD: Random Forest N=10	88.75±1.20	0.83±0.02	87.98±1.07	0.82±0.02
	AD: Random Forest N=25	89.41±0.72	0.86±0.02	88.79±0.80	0.85±0.02
	AD: Random Forest N=50	89.64±0.79	0.87±0.02	89.25±0.73	0.86±0.02
	AD: Random Forest N=75	89.74±0.88	0.87±0.02	89.38±0.82	0.87±0.02
	AD: Random Forest N=100	89.96±0.81	0.87±0.02	89.29±0.93	0.87±0.02
Con IG	*RB: K2*	86,73±0,74	0,81±0,03	86,84±0,54	0,81±0,02
	RB: TAN	87,71±0,33	0,72±0,03	87,54±0,51	0,73±0,03
	Naïve Bayes	27,80±1,93	0,76±0,03	28,90±2,54	0,71±0,04
	SVM: Polinomial	80,92±1,70	0,86±0,02	82,53±1,13	0,86±0,03
	SVM: Polinomial norm.	86,25±1,21	0,87±0,02	85,38±1,54	0,87±0,03
	SVM: Pearson VII	88,66±1,42	0,87±0,02	87,83±1,49	0,87±0,03
	SVM: RBF	79,20±1,67	0,86±0,01	78,03±1,56	0,85±0,02
	KNN K=1	80,31±2,06	0,72±0,03	79,61±1,97	0,71±0,02
	KNN K=2	83,51±1,42	0,78±0,03	83,24±1,77	0,76±0,03
	KNN K=3	79,90±1,95	0,80±0,03	79,01±1,79	0,79±0,03
	KNN K=4	81,04±1,73	0,81±0,03	80,42±2,07	0,80±0,03
	KNN K=5	78,58±1,67	0,82±0,03	77,90±1,88	0,81±0,03
	AD: J48	87,16±1,18	0,75±0,03	87,67±1,01	0,77±0,05
	AD: Random Forest N=10	90,08±0,65	0,86±0,02	89,88±0,98	0,87±0,03
	AD: Random Forest N=25	90,68±0,74	0,88±0,02	90,56±1,29	0,88±0,03
	AD: Random Forest N=50	90,71±1,09	0,89±0,02	90,55±1,09	0,89±0,03
	AD: Random Forest N=75	90,68±0,97	0,89±0,02	90,75±1,05	0,89±0,02
	AD: Random Forest N=100	90,62±0,94	0,89±0,02	90,84±0,99	0,89±0,02

5.2.9 **Marco primario**

La Tabla 5.11 recoge los resultados de la categorización del marco primario de los mensajes utilizando *Information Gain*. Se puede observar que la aplicación de esta técnica ofrece resultados inferiores en precisión cuando se utilizan N-gramas, en la medida en que tan solo 5 algoritmos mejoran sus puntuaciones mientras que son 13 los algoritmos que aumentan su precisión. En cambio, al utilizar Tokens, la precisión aumenta si se usa *Information Gain* en 10 casos, en 1 se mantiene y se reduce en otros 7.

La *Ganancia de Información* mejora el AUC, tanto utilizando N-gramas como Tokens. En el primero de los casos, 11 técnicas incrementan su área bajo la curva ROC, 5 se conserva y 2 empeoran; cuando se utilizan Tokens, 13 obtienen mejores resultados, 4 consiguen el mismo y solo uno de ellos retrocede.

El mayor beneficiado proporcionalmente son las máquinas de soporte vectorial utilizando *Pearson VII*, con aumentos muy notables en precisión, y algo menos en área bajo la curva ROC. Los algoritmos que más han mejorado en AUC son *K*-vecinos más próximos, aunque han perdido significativamente cuando emplean N-gramas. En cambio, los mayores retrocesos en precisión son para las máquinas de soporte vectorial con núcleo RBF.

En conjunto -utilizando y sin utilizar *Information Gain*- el algoritmo más eficaz son los árboles de decisión con *Random Forest* de 100 árboles utilizando Tokens e incorporando *Ganancia de Información*, que logra una precisión del 53,08 % y un área bajo la curva ROC de 0,8. En el entorno de estos resultados se encuentran otros 23 que son estadísticamente equivalentes. En este grupo se encuentran los árboles de decisión con *Random Forest* de más de 10 árboles y las redes bayesianas con núcleos K2 y TAN. En el grupo más retrasado, *Naïve Bayes* ocupa los resultados más pobres, con precisiones en torno al 20 % y áreas bajo la curva ROC inferiores a 0,7.

5.2.10 **Marco secundario**

Los resultados de la Tabla 5.12 muestran la categorización incorporando la aplicación de *Information Gain*. Tal como se aprecia, la mayoría de algoritmos obtienen peor precisión con *Ganancia de Información* que sin ella (16 frente a 2), tanto cuando se hace uso de N-gramas como de Tokens. Sin embargo, el área bajo la curva ROC mejora en términos similares: de los 18 algoritmos empleando N-gramas, 16 alcanzan mayor valor de AUC, mientras uno se mantiene y otro llega a valores menores. Cuando se utilizan Tokens, 15 obtienen mejores resultados, 2 los conservan y solo uno retrocede.

Tabla 5.11: Resultados comparados del uso de *Information Gain* para clasificar el marco primario en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	49,16±1,19	0,76±0,01	48,82±1,39	0,77±0,01
	RB: TAN	49,07±2,34	0,78±0,01	47,77±2,50	0,78±0,01
	Naïve Bayes	18,25±1,40	0,67±0,01	19,02±1,59	0,66±0,01
	SVM: Polinomial	41,21±1,27	0,75±0,01	41,99±1,68	0,75±0,01
	SVM: Polinomial norm.	46,34±1,41	0,78±0,01	46,46±2,69	0,78±0,01
	SVM: Pearson VII	33,59±6,24	0,73±0,02	35,00±6,11	0,74±0,02
	SVM: RBF	31,07±1,85	0,73±0,01	26,54±2,35	0,71±0,01
	KNN K=1	39,98±2,59	0,56±0,02	38,79±1,67	0,56±0,01
	KNN K=2	37,43±2,01	0,58±0,02	35,68±2,22	0,57±0,02
	KNN K=3	35,09±2,09	0,58±0,02	32,32±2,85	0,57±0,02
	KNN K=4	34,04±3,01	0,58±0,02	31,25±3,14	0,57±0,02
	KNN K=5	33,31±3,18	0,59±0,02	29,94±3,33	0,58±0,02
	AD: J48	37,53±2,07	0,62±0,02	39,43±1,94	0,64±0,01
	AD: Random Forest N=10	45,32±1,93	0,73±0,01	45,30±1,85	0,73±0,01
	AD: Random Forest N=25	49,79±2,58	0,77±0,02	49,80±1,83	0,77±0,01
	AD: Random Forest N=50	51,94±1,84	0,79±0,01	51,35±1,94	0,79±0,01
	AD: Random Forest N=75	52,11±1,51	0,79±0,01	52,32±1,65	0,79±0,01
	AD: Random Forest N=100	52,61±1,77	0,80±0,01	52,61±2,02	0,80±0,01
Con IG	*RB: K2*	49,63±1,59	0,77±0,01	48,31±1,80	0,78±0,01
	RB: TAN	48,77±2,39	0,78±0,01	48,14±2,35	0,78±0,01
	Naïve Bayes	20,64±0,97	0,69±0,01	19,14±1,61	0,68±0,02
	SVM: Polinomial	40,22±1,34	0,76±0,01	39,66±1,53	0,76±0,01
	SVM: Polinomial norm.	44,81±1,71	0,77±0,01	45,01±1,95	0,78±0,01
	SVM: Pearson VII	43,78±2,29	0,77±0,02	45,64±2,56	0,78±0,01
	SVM: RBF	22,75±2,48	0,71±0,01	20,00±2,07	0,70±0,01
	KNN K=1	37,78±2,57	0,60±0,02	38,77±1,62	0,61±0,01
	KNN K=2	34,57±1,61	0,62±0,01	35,94±1,75	0,63±0,01
	KNN K=3	31,41±1,13	0,62±0,01	32,58±1,46	0,64±0,01
	KNN K=4	30,53±0,98	0,63±0,01	31,44±1,36	0,65±0,01
	KNN K=5	29,55±1,51	0,64±0,01	30,64±1,94	0,66±0,01
	AD: J48	38,35±3,07	0,64±0,02	39,34±1,30	0,65±0,02
	AD: Random Forest N=10	45,54±1,21	0,73±0,02	45,60±1,71	0,74±0,01
	AD: Random Forest N=25	48,44±1,99	0,77±0,01	49,31±0,86	0,78±0,01
	AD: Random Forest N=50	50,88±1,80	0,79±0,01	51,44±2,02	0,79±0,01
	AD: Random Forest N=75	52,10±1,81	0,80±0,01	52,32±1,78	0,80±0,01
	AD: Random Forest N=100	52,23±1,86	0,80±0,02	53,08±1,85	0,80±0,01

Tabla 5.12: Resultados comparados del uso de *Information Gain* para clasificar el marco secundario en mensajes, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	87,83± 1,41	0,75±0,05	87,41± 1,09	0,77±0,04
	RB: TAN	75,19± 2,94	0,77±0,04	75,09± 2,84	0,77±0,04
	Naïve Bayes	11,38± 1,46	0,58±0,03	14,46± 1,42	0,62±0,04
	SVM: Polinomial	67,71± 5,41	0,73±0,04	76,52± 3,42	0,71±0,04
	SVM: Polinomial norm.	83,89± 3,55	0,76±0,03	84,50± 3,74	0,75±0,03
	SVM: Pearson VII	80,59±27,25	0,66±0,10	84,87±23,10	0,64±0,10
	SVM: RBF	64,85± 2,92	0,79±0,03	58,95± 3,82	0,77±0,04
	KNN K=1	87,64± 2,17	0,56±0,04	90,43± 0,62	0,52±0,03
	KNN K=2	83,87± 3,69	0,58±0,05	89,57± 1,01	0,53±0,03
	KNN K=3	80,50± 4,59	0,57±0,06	88,77± 1,32	0,53±0,03
	KNN K=4	77,50± 5,19	0,57±0,05	87,28± 1,81	0,54±0,03
	KNN K=5	75,15± 5,61	0,57±0,04	86,18± 2,09	0,55±0,03
	AD: J48	63,26± 3,69	0,61±0,04	64,70± 2,87	0,65±0,03
	AD: Random Forest N=10	85,64± 1,72	0,76±0,04	85,24± 1,05	0,77±0,04
	AD: Random Forest N=25	88,05± 1,48	0,78±0,04	87,47± 1,13	0,79±0,04
	AD: Random Forest N=50	88,96± 1,14	0,80±0,03	88,53± 1,15	0,80±0,03
	AD: Random Forest N=75	89,29± 1,07	0,80±0,03	89,23± 1,00	0,81±0,03
	AD: Random Forest N=100	89,31± 1,02	0,81±0,04	89,44± 0,92	0,81±0,03
Con IG	*RB: K2*	84,73±2,20	0,79±0,05	83,49±1,84	0,79±0,04
	RB: TAN	73,44±3,31	0,78±0,04	74,23±2,83	0,78±0,04
	Naïve Bayes	12,21±1,64	0,71±0,04	15,49±1,32	0,68±0,05
	SVM: Polinomial	62,06±2,42	0,75±0,02	57,14±2,58	0,76±0,03
	SVM: Polinomial norm.	70,78±3,42	0,77±0,02	67,03±3,27	0,77±0,03
	SVM: Pearson VII	61,99±7,39	0,77±0,03	69,04±6,24	0,78±0,03
	SVM: RBF	46,43±4,71	0,77±0,04	38,69±5,66	0,76±0,04
	KNN K=1	76,79±1,55	0,58±0,03	74,68±1,80	0,61±0,03
	KNN K=2	68,94±2,33	0,61±0,04	64,82±1,66	0,64±0,04
	KNN K=3	63,23±2,12	0,62±0,03	58,14±1,93	0,65±0,04
	KNN K=4	60,16±2,43	0,63±0,04	53,65±2,01	0,65±0,03
	KNN K=5	56,70±2,68	0,64±0,04	50,05±2,45	0,65±0,03
	AD: J48	66,00±2,30	0,65±0,04	67,14±2,47	0,66±0,05
	AD: Random Forest N=10	83,70±1,73	0,77±0,03	83,26±1,62	0,78±0,04
	AD: Random Forest N=25	85,87±1,72	0,80±0,03	85,07±1,45	0,80±0,03
	AD: Random Forest N=50	86,87±1,47	0,81±0,03	86,14±1,86	0,81±0,04
	AD: Random Forest N=75	87,33±1,58	0,81±0,03	86,49±1,99	0,81±0,04
	AD: Random Forest N=100	87,51±1,80	0,81±0,03	86,75±1,99	0,81±0,04

Las máquinas de soporte vectorial y los K-vecinos más próximos sufren reducciones notables porcentualmente en precisión mientras que mejoran en área bajo la curva ROC (excepto SVM con RBF, que obtiene peores resultados en am-

bos). El resto de técnicas -redes bayesianas y árboles de decisión- no varían significativamente salvo *Naïve Bayes* que avanza notablemente y en algo menos lo hacen los árboles de decisión con la implementación *J48*.

Los mejores resultados los logran los árboles de decisión con bosques aleatorios de 100 árboles sin emplear *Information Gain* y utilizando Tokens, que alcanzan una precisión del 89,44 % y un área bajo la cura ROC igual a 0,81. Este algoritmo no es el que más precisión tiene, pero aúna una buena precisión junto con un valor elevado del área bajo la curva ROC. El algoritmo con mayor precisión (90,43 %) es K-vecinos próximos con K=1 empleando Tokens pero no *Ganancia de Información* que, en cambio, ofrece un AUC de apenas 0,52.

En el entorno del mejor algoritmo se encuentran otros 17, que en su mayoría son árboles de decisión con bosques aleatorios, además de las redes bayesianas con el núcleo K2 empleando N-gramas y sin aplicar *Ganancia de Información*. Los *Naïve Bayes* en cualquiera de las cuatro modalidades empleadas ofrecen los resultados más débiles del global, con precisiones por debajo del 15 % y valores de AUC entre 0,58 y 0,71.

5.2.11 **Análisis del modelo de mensajes**

Los resultados comparados de las mejores técnicas de clasificación de mensajes -con y sin *Ganancia de Información*- se muestran en la Tabla 5.13. Como se puede observar, la aplicación de esta técnica supone un avance cuantitativo y cualitativo.

Así, se han superado de algunas de las deficiencias que presentaba el modelo básico de clasificación de los mensajes. Precisamente las clasificación que ofrecían resultados más pobres (grado de debate y marco primario) han mejorado mediante la incorporación de *Information Gain*. No obstante, otras clasificaciones obtienen peores cifras en precisión o en área bajo la curva ROC, como son la función movilizadora, el nivel de meta-comunicación, el grado de contribución y el marco secundario.

Realizando una selección de las mejores técnicas, el modelo mejorado para la clasificación de mensajes es el formado por cada uno de los mejores métodos de clasificación, y que se detalla en la Tabla 5.14. Este modelo incorpora árboles de decisión para 8 de las 9 categorías, mientras que las redes bayesianas con el núclero *TAN* completa el conjunto de técnicas. Cinco de los 10 clasificadores utilizan Tokens en el modelo de espacio vectorial, mientras que otros 5 emplean N-gramas. Igualmente, el modelo incluye 7 algoritmos sobre los que se aplica *Ganancia de Información*, mientras que 3 de ellos no lo tienen incorporado.

Tabla 5.13: Comparativa de mejores resultados de categorización de mensajes, en términos de precisión (%) y AUC.

		Algoritmo	VSM	Prec.(%)	AUC
Sin IG	Func. informativa	*AD: Random Forest N=50*	Tokens	97,21	0,98
	Func. movilizadora	*AD: Random Forest N=100*	N-gramas	89,06	0,79
	Meta-comunicación	*AD: Random Forest N=100*	Tokens	78,96	0,8
	Campaña negativa	*AD: Random Forest N=100*	N-gramas	89,96	0,87
	Nivel personalización	*AD: Random Forest N=100*	Tokens	79,33	0,87
	Grado contribución	*RB: TAN*	N-gramas	82,34	0,81
	Grado debate	*AD: Random Forest N=100*	Tokens	60,67	0,78
	Referenc. contenidos	*AD: Random Forest N=75*	Tokens	91,99	0,87
	Marco primario	*AD: Random Forest N=100*	N-gramas	52,61	0,8
	Marco secundario	*AD: Random Forest N=100*	Tokens	89,44	0,81
Con IG	Func. informativa	*AD: Random Forest N=75*	Tokens	97,53	0,98
	Func. movilizadora	*RB: K2*	N-gramas	87,82	0,75
	Meta-comunicación	*AD: Random Forest N=100*	N-gramas	79,18	0,81
	Campaña negativa	*AD: Random Forest N=100*	Tokens	90,84	0,89
	Nivel personalización	*AD: Random Forest N=100*	Tokens	80,86	0,89
	Grado contribución	*AD: Random Forest N=75*	Tokens	80,53	0,82
	Grado debate	*AD: Random Forest N=100*	N-gramas	65,13	0,78
	Referenc. contenidos	*AD: Random Forest N=100*	N-gramas	92,53	0,9
	Marco primario	*AD: Random Forest N=100*	Tokens	53,08	0,8
	Marco secundario	*AD: Random Forest N=100*	N-gramas	87,51	0,81

5.3 Relaciones entre las variables del modelo de acción comunicativa

Tal como se ha descrito en los capítulos 2 y 3, los modelos de representación de la acción comunicativa de las conversaciones y de los mensajes que las componen se estructura en varias dimensiones, cada una de las cuales establece una serie de variables que permiten operacionalizar su observación y análisis.

De este modo, para las conversaciones políticas, las dimensiones que se establecen son el análisis funcional, el análisis argumentativo y la estrategia comunicativa de dicha acción, en nuestro caso estudiada desde la perspectiva del análisis de marcos. El análisis funcional incluiría la función informativa y la función movilizadora. Por su parte, análisis argumentativo abarcaría el nivel de meta-comunicación, el nivel de personalización y el nivel de campaña negativa. Por último la estrategia comunicativa incluye el marco primario y el marco secundario.

Por su parte, los mensajes, además de las dimensiones de las conversaciones,

Tabla 5.14: Modelo mejorado de clasificación de mensajes, en términos de precisión (%) y AUC.

	Algoritmo	VSM	IG	Prec.(%)	AUC
Función informativa	*AD: Random Forest N=75*	*Tokens*	*Sí*	97,53	0,98
Función movilizadora	*AD: Random Forest N=100*	*N-gramas*	*No*	89,06	0,79
Meta-comunicación	*AD: Random Forest N=100*	*N-gramas*	i	79,18	0,81
Campaña negativa	*AD: Random Forest N=100*	*Tokens*	*Sí*	90,84	0,89
Nivel personalización	*AD: Random Forest N=100*	*Tokens*	*Sí*	80,86	0,89
Grado contribución	*RB: TAN*	*N-gramas*	*No*	82,34	0,81
Grado debate	*AD: Random Forest N=100*	*N-gramas*	*Sí*	65,13	0,78
Referenc. contenidos	*AD: Random Forest N=100*	*N-gramas*	*Sí*	92,53	0,9
Marco primario	*AD: Random Forest N=100*	*Tokens*	*Sí*	53,08	0,8
Marco secundario	*AD: Random Forest N=100*	*Tokens*	*No*	89,44	0,81

añaden otra correspondiente al nivel deliberativo propio de cada una de las interacciones, a saber: el grado de contribución, el grado de debate político y la referencialidad de los contenidos.

Las variables de cada una de las dimensiones del modelo son independientes entre sí, y de esta forma se han tratado en el proceso de etiquetado, tanto para los mensajes como para las conversaciones. Por todo ello, es posible incluirlas en los modelos de clasificación que se testean para su validación, y así poder crear un modelo sofisticado.

5.3.1 Análisis de la *Ganancia de Información* en las conversaciones políticas

Se ha realizado un análisis de *Ganancia de Información* con el propósito de identificar posibles relaciones entre las variables utilizadas para las clasificaciones de conversaciones políticas. Como en el caso de los mensajes, este análisis se ha efectuado considerando las relaciones directas entre las variables que representan las categorías etiquetadas de las conversaciones, con respecto a cada una de las demás. Por tanto, este análisis no describe las posibles relaciones cruzadas entre todas las categorías entre sí, sino una a una.

A continuación se describen las relaciones que se han identificado a partir de este análisis en cada una de las variables de las conversaciones.

Función informativa . Los resultados del análisis de ganancia de información, presentados en la Tabla 5.15, muestran la existencia de correlaciones en la categoría función informativa, que está influenciada de manera significativa

por el marco secundario, el marco primario y el nivel de personalización de la conversación.

Tabla 5.15: Análisis de ganancia de información de la función informativa de las conversaciones políticas.

Valor	Clase
0,3036	Marco secundario
0,178	Marco primario
0,1184	Nivel de personalización
0,0955	Nivel de campaña negativa
0,0905	Nivel de meta-comunicación
0,0553	Función movilizadora

Función movilizadora . Tal como muestra la Tabla 5.16, la categoría función movilizadora de una conversación, analizada de manera individual, tiene notables correlaciones con su marco secundario, su marco primario, su nivel de personalización y su nivel de meta-comunicación.

Tabla 5.16: Análisis de ganancia de información de la función movilizadora de las conversaciones políticas.

Valor	Clase
0,2676	Marco secundario
0,2431	Marco primario
0,1922	Nivel de personalización
0,1631	Nivel de meta-comunicación
0,0553	Función informativa
0,0307	Nivel de campaña negativa

Nivel de personalización . La Ganancias de Información, tal como aparece en la Tabla 5.17 muestra que el nivel de personalización tiene una notable relación con el resto de variables, y principalmente con el marco primario y el nivel de campaña negativa, como puede observarse en la Tabla 5.17.

Nivel de meta-comunicación Los datos de la Tabla 5.18 apuntan a que existe una intensa relación entre el nivel de meta-comunicación de los mensajes y el resto de las variables etiquetadas de los mensajes, con especial intensidad el marco primario y el marco secundario.

Tabla 5.17: Análisis de ganancia de información del nivel de personalización de las conversaciones políticas.

Valor	Clase
0,378	Marco primario
0,241	Nivel de campaña negativa
0,195	Marco secundario
0,192	Función movilizadora
0,185	Nivel de meta-comunicación
0,118	Función informativa

Tabla 5.18: Análisis de ganancia de información del nivel de meta-comunicación de las conversaciones políticas.

Valor	Clase
0,5292	Marco secundario
0,3135	Marco primario
0,1851	Nivel de personalización
0,1631	Función movilizadora
0,1041	Nivel de campaña negativa
0,0905	Función informativa

Nivel de campaña negativa El análisis de ganancia de información realizado sobre las conversaciones políticas muestra que el nivel de campaña negativa está influenciado estadísticamente por el nivel de personalización, el marco primario y el nivel de meta-comunicación etiquetados de las conversaciones, tal como muestra la Tabla 5.19.

Tabla 5.19: Análisis de ganancia de información del nivel de campaña negativa de las conversaciones políticas.

Valor	Clase
0,241	Nivel de personalización
0,155	Marco primario
0,1041	Nivel de meta-comunicación
0,0955	Función informativa
0,0572	Marco secundario
0,0307	Función movilizadora

Marco primario Según muestra la Tabla 5.20, el análisis de *Ganancia de Información* señala que el marco primario de las conversaciones tiene una gran relación con el marco secundario,el nivel de personalización y el nivel de meta-comunicación, y algo menos con el resto de características etiquetadas en los mensajes de la conversación.

Tabla 5.20: Análisis de ganancia de información del marco primario de las conversaciones políticas.

Valor	Clase
0,853	Marco secundario
0,378	Nivel de personalización
0,313	Nivel de meta-comunicación
0,243	Función movilizadora
0,178	Función informativa
0,155	Nivel de campaña negativa

Marco secundario En la Tabla 5.21 se muestran las principales relaciones de *Ganancia de Información* identificadas tras el proceso de etiquetado. Estos resultados muestran que el marco secundario de una conversación está relacionado, según orden de importancia decreciente, con el marco primario, el nivel de meta-comunicación, la función informativa, la función movilizadora y el nivel de personalización. No se aprecia, en cambio, una relación procedente del nivel de campaña negativa.

Tabla 5.21: Análisis de ganancia de información del marco secundario de las conversaciones políticas.

Valor	Clase
0,8532	Marco primario
0,5292	Nivel de meta-comunicación
0,3036	Función informativa
0,2676	Función movilizadora
0,195	Nivel de personalización
0,0572	Nivel de campaña negativa

5.3.2 Discusión de resultados

5.3.2.1 Función informativa

Una vez realizados los análisis incorporando *Ganancia de Información*, que se muestra en la Tabla 5.22, se comprueba que el uso de esta técnica mejora los niveles de precisión y del área bajo la curva ROC en la mayoría de los casos. Respecto a la precisión, en solo 6 algoritmos presentan peores resultados al utilizar *Information Gain* (4 de ellos usan N-gramas y los otros 2 utilizan Tokens). En relación al valor AUC, se repite la misma tónica, pero en este caso apenas 2 algoritmos presentan igual resultado al aplicar *Ganancia de Información*, mientras que uno solo (las máquinas de soporte vectorial con núcleo normalizado) obtiene peor resultado.

La mejor categorización de la función informativa es la alcanzada con los árboles de decisión con bosques aleatorios compuestos por 75 árboles (N=75), que consigue una precisión del 68,57 % y un área bajo la curva ROC de 0,83. Esto supone una pequeña mejora con respecto al modelo básico sin usar IG, de 6 puntos en precisión (62,32 % sin *Information Gain*), y un pequeño aumento en el AUC (0,79 sin Information Gain).

La mejor precisión se logra con el uso de *KNN* (con K=1), con un 70,89 % y un AUC de 0,76 cuando se usan Tokens, y una precisión del 69,64 % y un AUC de 0,76 utilizando N-gramas. Por otro lado, los mejores valores del ára bajo la curva ROC los obtenemos utilizando redes bayesianas. Con *Naïve Bayes u*sando Tokens, llegamos a un 0,85, pero la precisión es tan solo del 66,07 %, mientras que con redes bayesianas y el núcleo *TAN* logramos un AUC=0,84 y un 64,82 % de precisión.

5.3.2.2 Función movilizadora

La Tabla 5.23 muestra los resultados de categorizar la función movilizadora de manera comparativa entre la aplicación o no de la técnica de *Ganancia de Información*. En general, el uso de esta técnica mejora los valores alcanzados sin ella tanto en lo que respecta a la precisión del modelo como en cuanto al área descrito por la curva ROC, con una sola excepción, las máquinas de soporte vectorial (SVM) con el núcleo *Pearson VII* y N-gramas, que redujeron su precisión. Los algoritmos más beneficiados son *KNN* y AD con *J48*, tanto en N-gramas como cuando utilizan Tokens, que logran significativos avances.

Los mejores resultados en precisión son los alcanzados por un grupo de algoritmos. El mejor de ellos son las redes bayesianas con el núcleo *K2* empleando N-gramas, que logra una precisión del 83,57 % y un área bajo la curva ROC igual

Tabla 5.22: Resultados comparados del uso de *Information Gain* para clasificar la función informativa en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	58,39±16,37	0,78±0,15	58,39±17,40	0,73±0,14
	RB: TAN	63,57±15,80	0,76±0,13	54,82±21,11	0,77±0,16
	Naïve Bayes	59,46±14,19	0,70±0,13	58,39±15,27	0,71±0,12
	SVM: Polinomial	67,32±11,76	0,77±0,09	62,14±11,41	0,78±0,09
	SVM: Polinomial norm.	61,07±15,77	0,72±0,13	68,57±11,85	0,73±0,12
	SVM: Pearson VII	48,93±22,86	0,69±0,14	63,57±13,42	0,69±0,16
	SVM: RBF	62,32±17,40	0,68±0,15	56,07±18,54	0,72±0,17
	KNN K=1	44,82±22,61	0,59±0,13	54,64±20,22	0,68±0,15
	KNN K=2	54,64±20,22	0,65±0,09	55,89±19,42	0,73±0,16
	KNN K=3	55,71±17,79	0,64±0,09	51,96±18,20	0,73±0,17
	KNN K=4	49,64±13,78	0,62±0,11	51,96±18,20	0,74±0,12
	KNN K=5	43,39±19,16	0,63±0,15	43,93±17,58	0,73±0,11
	AD: J48	55,18±15,28	0,67±0,14	42,86±15,41	0,60±0,13
	AD: Random Forest N=10	65,00±17,48	0,75±0,18	54,82±16,50	0,68±0,10
	AD: Random Forest N=25	59,82±19,09	0,74±0,20	54,82±14,24	0,73±0,09
	AD: Random Forest N=50	59,82±17,17	0,79±0,13	62,32±14,10	0,77±0,08
	AD: Random Forest N=75	57,32±13,00	0,76±0,13	59,82±13,81	0,78±0,09
	AD: Random Forest N=100	58,57±14,11	0,77±0,13	62,32±15,28	0,79±0,09
Con IG	*RB: K2*	60,71± 9,30	0,80±0,10	64,82±16,19	0,81±0,13
	RB: TAN	62,14±16,41	0,82±0,14	64,82±16,19	0,84±0,12
	Naïve Bayes	64,64±13,95	0,83±0,11	66,07±18,46	0,85±0,13
	SVM: Polinomial	56,96±10,39	0,76±0,14	66,07±15,38	0,82±0,12
	SVM: Polinomial norm.	62,14± 9,77	0,81±0,11	63,39±12,12	0,83±0,10
	SVM: Pearson VII	60,71±12,49	0,82±0,08	61,96±11,91	0,83±0,10
	SVM: RBF	56,96±14,57	0,77±0,10	59,64±12,49	0,77±0,11
	KNN K=1	69,64±10,41	0,76±0,08	70,89±13,18	0,76±0,10
	KNN K=2	63,39±13,48	0,81±0,09	65,89±10,01	0,80±0,10
	KNN K=3	62,14±19,32	0,78±0,12	62,14±19,32	0,76±0,12
	KNN K=4	58,39±21,82	0,76±0,13	59,64±17,17	0,74±0,10
	KNN K=5	58,39±20,17	0,76±0,12	53,39±17,28	0,73±0,11
	AD: J48	52,14±19,83	0,68±0,14	52,14±14,82	0,65±0,10
	AD: Random Forest N=10	60,89±15,78	0,78±0,11	64,64± 9,52	0,81±0,13
	AD: Random Forest N=25	62,14±12,84	0,80±0,09	67,32±15,57	0,81±0,14
	AD: Random Forest N=50	62,14± 9,77	0,82±0,07	68,57±13,24	0,82±0,14
	AD: Random Forest N=75	59,64±11,01	0,81±0,08	68,57±11,85	0,83±0,13
	AD: Random Forest N=100	60,89±12,04	0,81±0,08	66,07±11,51	0,83±0,11

a 0,83. En el mismo entorno se encuentran *Naïve Bayes* con tokens, que ofrece una precisión del 80,89 % y un AUC de valor igual a 0,85; y las redes bayesianas con el núcleo TAN, que llega al 79,82 % de precisión y a un 0,86 de área bajo la curva ROC. Los K-vecinos más cercanos utilizando N-gramas, con valores de K entre 2 y 5 obtienen los resultados más pobres en términos de precisión y AUC.

5.3.2.3 **Nivel de meta-comunicación**

De la aplicación de *Ganancia de Información* a la clasificación del nivel de metacomunicación se desprenden los resultados que se muestran en la Tabla 5.24. Como se puede observar, esto ha supuesto una mejora en los resultados de todos los algoritmos, tanto en precisión como un incremento en el área bajo la curva ROC. La familia de los K-vecinos más próximos ha sido la que más beneficiada ha sido, tanto en precisión como en el valor de AUC, mientras que los árboles de decisión con bosques aleatorios y las redes bayesianas simples han incrementado sustancialmente en el valor del área bajo la curva.

Los mejores resultados son los alcanzados con la aplicación de *Information Gain* sobre los árboles de decisión con bosques aleatorios de 100 árboles, logrando una precisión del 70,71 % y un área bajo la curva ROC igual a 0,83. Así mismo, en el entorno de estos resultados se encuentran 6 de los algoritmos utilizados, como K-vecinos más próximos y valor de K=1, que logra una precisión del 69,29 % y un valor de AUC = 0,78, tanto usando N-gramas como Tokens. Mediante el uso de árboles de decisión con bosques aleatorios (con N=50, N=75) y empleando N-gramas se obtiene una precisión del 68,21 % y un área bajo la curva ROC de 0,85, de manera similar a los casos en que se usan 100 árboles y N-gramas, y 25 árboles con Tokens.

Por el contrario, si no se aplica IG se mantienen resultados más pobres, tanto en precisión como en la obtención del área bajo la curva ROC. Así, K-vecinos más próximos empleando N-gramas, con K=3 apenas llega al 10 % de precisión y un AUC igual a 0,54. La misma variante con K=35 apenas sube hasta el 12,68 % de precisión y un AUC de valor 0,57.

5.3.2.4 **Nivel de personalización**

Teniendo en cuenta los resultados de aplicar los distintos algoritmos, utilizando tanto N-gramas como términos, y comparando los obtenidos después de aplicar *Information Gain* (véase la Tabla 5.25), el mejor resultado es el alcanzado por los árboles de decisión. Para ello se ha utilizado la variante de bosques aleatorios, con 50 árboles ($N = 50$) y utilizando Tokens, tras haber realizado un análisis de

Tabla 5.23: Resultados comparados del uso de *Information Gain* para clasificar la función movilizadora en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	49,64±13,78	0,60±0,15	49,46±12,71	0,59±0,18
	RB: TAN	56,96±11,95	0,62±0,19	50,71±13,37	0,58±0,15
	Naïve Bayes	51,96±12,56	0,50±0,11	36,79±20,05	0,51±0,20
	SVM: Polinomial	50,71± 8,63	0,56±0,10	54,46±13,21	0,65±0,15
	SVM: Polinomial norm.	51,96± 9,39	0,62±0,06	50,71±12,00	0,62±0,09
	SVM: Pearson VII	42,86±16,50	0,53±0,13	48,04±13,87	0,57±0,13
	SVM: RBF	49,46±12,71	0,59±0,12	54,64±15,34	0,62±0,15
	KNN K=1	27,68±12,69	0,51±0,13	29,11±18,63	0,53±0,14
	KNN K=2	26,61± 9,24	0,51±0,10	30,36±15,73	0,52±0,12
	KNN K=3	21,43±10,14	0,52±0,13	30,18±18,64	0,51±0,17
	KNN K=4	22,68±11,33	0,51±0,15	22,68±19,27	0,53±0,19
	KNN K=5	21,43±10,14	0,53±0,11	26,43±13,57	0,52±0,20
	AD: J48	37,14±18,80	0,46±0,18	45,36±22,65	0,59±0,19
	AD: Random Forest N=10	50,36±16,10	0,56±0,12	51,96±11,09	0,53±0,16
	AD: Random Forest N=25	51,96± 9,39	0,56±0,09	49,46±12,71	0,62±0,14
	AD: Random Forest N=50	53,21± 9,92	0,59±0,07	45,71±11,14	0,62±0,11
	AD: Random Forest N=75	50,71±10,45	0,57±0,10	48,21±11,87	0,64±0,11
	AD: Random Forest N=100	51,96± 9,39	0,60±0,11	50,71±10,45	0,63±0,11
Con IG	*RB: K2*	60,89±16,85	0,81±0,15	60,89±17,85	0,79±0,16
	RB: TAN	66,07±15,38	0,76±0,21	66,07±11,51	0,75±0,20
	Naïve Bayes	50,54±11,26	0,81±0,11	61,79±15,27	0,81±0,13
	SVM: Polinomial	63,57±12,06	0,77±0,16	64,82±12,57	0,77±0,16
	SVM: Polinomial norm.	66,07±12,93	0,78±0,17	66,07±12,93	0,78±0,16
	SVM: Pearson VII	56,07±16,56	0,79±0,17	59,82±17,17	0,79±0,18
	SVM: RBF	58,39±16,37	0,77±0,18	58,39±16,37	0,77±0,18
	KNN K=1	70,00±22,97	0,76±0,19	70,00±22,97	0,76±0,19
	KNN K=2	71,25±17,73	0,74±0,18	69,82±15,60	0,73±0,17
	KNN K=3	62,32±18,37	0,73±0,18	62,32±16,38	0,73±0,17
	KNN K=4	62,32±15,28	0,73±0,17	59,64±11,01	0,72±0,17
	KNN K=5	62,32±17,40	0,73±0,18	61,07±16,83	0,71±0,18
	AD: J48	58,57±14,11	0,77±0,10	58,57±16,38	0,76±0,13
	AD: Random Forest N=10	64,82±13,88	0,79±0,15	64,64±12,65	0,77±0,16
	AD: Random Forest N=25	63,75±19,05	0,77±0,17	64,64±13,95	0,77±0,14
	AD: Random Forest N=50	67,50±15,81	0,79±0,16	72,50±15,37	0,78±0,15
	AD: Random Forest N=75	67,50±15,81	0,79±0,16	69,82±14,44	0,77±0,14
	AD: Random Forest N=100	66,25±16,72	0,79±0,17	65,89±11,62	0,77±0,15

Tabla 5.24: Resultados comparados del uso de *Information Gain* para clasificar el nivel de meta-comunicación en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	38,04±19,62	0,60±0,19	43,21±22,58	0,62±0,18
	RB: TAN	39,29±13,81	0,61±0,16	44,46±19,26	0,63±0,16
	Naïve Bayes	45,54±15,62	0,61±0,13	35,54±14,40	0,57±0,13
	SVM: Polinomial	36,79±18,23	0,65±0,17	39,29±16,13	0,62±0,11
	SVM: Polinomial norm.	43,04±11,95	0,63±0,17	45,89±18,09	0,64±0,14
	SVM: Pearson VII	31,61±14,61	0,58±0,15	43,04±16,78	0,59±0,14
	SVM: RBF	37,86±14,13	0,59±0,13	43,21±21,80	0,63±0,20
	KNN K=1	17,68±15,76	0,54±0,08	25,36±14,48	0,51±0,11
	KNN K=2	17,68±10,46	0,55±0,07	27,86±16,44	0,51±0,11
	KNN K=3	10,00±11,49	0,54±0,09	25,36±11,84	0,56±0,11
	KNN K=4	15,36±13,26	0,57±0,08	25,36±14,48	0,57±0,10
	KNN K=5	12,68±13,19	0,57±0,08	29,29±14,95	0,60±0,10
	AD: J48	37,68±18,37	0,56±0,16	40,54±19,36	0,60±0,15
	AD: Random Forest N=10	33,93±14,21	0,55±0,13	48,21±18,69	0,66±0,14
	AD: Random Forest N=25	39,29±11,01	0,59±0,07	41,96±20,91	0,67±0,18
	AD: Random Forest N=50	35,18±16,19	0,57±0,04	44,46±22,58	0,68±0,17
	AD: Random Forest N=75	35,18±15,08	0,58±0,08	43,39±24,70	0,65±0,18
	AD: Random Forest N=100	38,93±18,78	0,61±0,11	39,29±19,08	0,65±0,19
Con IG	*RB: K2*	50,54±11,26	0,73±0,12	50,54±18,31	0,72±0,14
	RB: TAN	56,96±14,57	0,75±0,13	53,21±14,23	0,73±0,13
	Naïve Bayes	50,89±13,10	0,74±0,08	49,46± 9,60	0,73±0,09
	SVM: Polinomial	61,96±11,91	0,82±0,10	63,21±13,91	0,83±0,12
	SVM: Polinomial norm.	63,21±12,60	0,83±0,12	63,21±11,14	0,83±0,12
	SVM: Pearson VII	60,54±21,34	0,83±0,15	59,29±21,57	0,82±0,15
	SVM: RBF	51,79±14,51	0,74±0,12	51,79±14,51	0,74±0,12
	KNN K=1	69,29±19,77	0,78±0,14	69,29±19,77	0,78±0,14
	KNN K=2	51,79± 8,46	0,81±0,14	55,54±12,75	0,82±0,13
	KNN K=3	54,29±11,14	0,77±0,09	58,04±12,63	0,76±0,12
	KNN K=4	56,79±14,06	0,75±0,10	59,29±16,03	0,76±0,13
	KNN K=5	43,04±16,78	0,73±0,11	48,04±20,87	0,75±0,14
	AD: J48	58,04±22,51	0,76±0,16	55,36±20,88	0,71±0,16
	AD: Random Forest N=10	64,29±13,98	0,83±0,11	61,96±26,41	0,78±0,18
	AD: Random Forest N=25	63,21±13,91	0,85±0,10	68,21±16,19	0,85±0,12
	AD: Random Forest N=50	68,21±11,10	0,85±0,11	64,46±17,65	0,84±0,14
	AD: Random Forest N=75	68,21±13,88	0,85±0,12	66,96±19,94	0,82±0,15
	AD: Random Forest N=100	68,21±15,08	0,84±0,11	70,71±18,10	0,83±0,15

ganancia de información, y llegando a una precisión del 72,5 % y un área bajo la curva ROC con valor igual a 0,78.

Comparando los resultados frente al modelo básico sin aplicación de IG, se aprecia una notable mejora. En el primer caso, superando el mejor resultado de precisión sin IG, que era de un 54,46 % y un área ROC de 0,65 (mediante *SVM* y núcleo polinomial usando términos).

En cuanto al área bajo la curva ROC, los mejores resultados son los que facilita Naïve Bayes usando tokens, con un 0,81, aunque la precisión del modelo es bastante inferior (61,79 %). De manera similar, las redes bayesianas con núcleo K2 también alcanzan un AUC = 0,81, pero con una precisión de tan solo el 60,89 %.

En términos comparados, el uso de *Information Gain* ha supuesto una mejora en la precisión de todos los modelos o algoritmos excepto en uno (*Supporting Vector Machines* con *RBF*, cuyo resultado ha empeorado), tanto empleando N-gramas como Tokens. No obstante, el área bajo la curva ROC ha aumentado en todos los algoritmos por la aplicación de esta técnica.

5.3.2.5 **Nivel de campaña negativa**

Los resultados comparados del uso de la *Ganancia de Información* para la clasificación de la campaña negativa se muestran en la Tabla 5.26.

El uso de esta técnica permite obtener mejores resultados en el área bajo la curva ROC en todos los algoritmos. En sentido contrario, su aplicación repercute negativamente en la precisión en 14 de los 18 algoritmos, mejorando dicha precisión en otros 4.

Los únicos algoritmos que salen favorecidos en todos los casos (con N-gramas y con Tokens) y tanto en precisión como en área bajo la curva ROC son los K-vecinos más próximos con valores de K desde 1 a 4 cuando utilizan N-gramas. En cambio las redes bayesianas en cualquiera de sus variantes, o las máquinas de soporte vectorial experimientan los mayores recortes. En términos comparativos, los retrocesos son mayores cuando se utilizan N-gramas que cuando se usan Tokens. Estos algoritmos ofrecen unos buenos resultados, tanto en precisión como en AUC.

Otro conjunto de algoritmos alcanzan precisiones muy altas, por encima del 85 %, pero en cambio el valor del área bajo la curva ROC no supera 0,6. Estos algoritmos no emplean *Ganancia de Información*, y son los formados por las máquinas de soporte vectorial con núcleos *Pearson VII* y polinomial normalizado, además de los árboles de decisión con bosques aleatorios y con 10 ó 50 árboles.

Tabla 5.25: Resultados comparados del uso de *Information Gain* para clasificar el nivel de personalización en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	49,64±13,78	0,60±0,15	49,46±12,71	0,59±0,18
	RB: TAN	56,96±11,95	0,62±0,19	50,71±13,37	0,58±0,15
	Naïve Bayes	51,96±12,56	0,50±0,11	36,79±20,05	0,51±0,20
	SVM: Polinomial	50,71± 8,63	0,56±0,10	54,46±13,21	0,65±0,15
	SVM: Polinomial norm.	51,96± 9,39	0,62±0,06	50,71±12,00	0,62±0,09
	SVM: Pearson VII	42,86±16,50	0,53±0,13	48,04±13,87	0,57±0,13
	SVM: RBF	49,46±12,71	0,59±0,12	54,64±15,34	0,62±0,15
	KNN K=1	27,68±12,69	0,51±0,13	29,11±18,63	0,53±0,14
	KNN K=2	26,61± 9,24	0,51±0,10	30,36±15,73	0,52±0,12
	KNN K=3	21,43±10,14	0,52±0,13	30,18±18,64	0,51±0,17
	KNN K=4	22,68±11,33	0,51±0,15	22,68±19,27	0,53±0,19
	KNN K=5	21,43±10,14	0,53±0,11	26,43±13,57	0,52±0,20
	AD: J48	37,14±18,80	0,46±0,18	45,36±22,65	0,59±0,19
	AD: Rand.Forest N=10	50,36±16,10	0,56±0,12	51,96±11,09	0,53±0,16
	AD: Rand.Forest N=25	51,96± 9,39	0,56±0,09	49,46±12,71	0,62±0,14
	AD: Rand.Forest N=50	53,21± 9,92	0,59±0,07	45,71±11,14	0,62±0,11
	AD: Rand.Forest N=75	50,71±10,45	0,57±0,10	48,21±11,87	0,64±0,11
	AD: Rand.Forest N=100	51,96± 9,39	0,60±0,11	50,71±10,45	0,63±0,11
Con IG	*RB: K2*	60,89±16,85	0,81±0,15	60,89±17,85	0,79±0,16
	RB: TAN	66,07±15,38	0,76±0,21	66,07±11,51	0,75±0,20
	Naïve Bayes	50,54±11,26	0,81±0,11	61,79±15,27	0,81±0,13
	SVM: Polinomial	63,57±12,06	0,77±0,16	64,82±12,57	0,77±0,16
	SVM: Polinomial norm.	66,07±12,93	0,78±0,17	66,07±12,93	0,78±0,16
	SVM: Pearson VII	56,07±16,56	0,79±0,17	59,82±17,17	0,79±0,18
	SVM: RBF	58,39±16,37	0,77±0,18	58,39±16,37	0,77±0,18
	KNN K=1	70,00±22,97	0,76±0,19	70,00±22,97	0,76±0,19
	KNN K=2	71,25±17,73	0,74±0,18	69,82±15,60	0,73±0,17
	KNN K=3	62,32±18,37	0,73±0,18	62,32±16,38	0,73±0,17
	KNN K=4	62,32±15,28	0,73±0,17	59,64±11,01	0,72±0,17
	KNN K=5	62,32±17,40	0,73±0,18	61,07±16,83	0,71±0,18
	AD: J48	58,57±14,11	0,77±0,10	58,57±16,38	0,76±0,13
	AD: Rand.Forest N=10	64,82±13,88	0,79±0,15	64,64±12,65	0,77±0,16
	AD: Rand.Forest N=25	63,75±19,05	0,77±0,17	64,64±13,95	0,77±0,14
	AD: Rand.Forest N=50	67,50±15,81	0,79±0,16	72,50±15,37	0,78±0,15
	AD: Rand.Forest N=75	67,50±15,81	0,79±0,16	69,82±14,44	0,77±0,14
	AD: Rand.Forest N=100	66,25±16,72	0,79±0,17	65,89±11,62	0,77±0,15

Dentro del uso de *Information Gain*, el uso de N-gramas o términos está parejo a la hora de obtener mayor precisión (cada uno de ellos obtiene mejores a partes iguales), por lo que el empleo de una u otra técnica de tratamiento de textos no supone diferencias.

Los mejores resultados en precisión son obtenidos sin IG, mediante máquinas de soporte vectorial con el núcleo *Pearson VII* y N-gramas, que llegan a un 88,75 %, aunque tan solo ofrecen un AUC igual a 0,36. Por su parte, los mejores resultados en AUC los logra K-vecinos más próximos con K=1 empleando Tokens y sin aplicar *Ganancia de Información*. De este modo alcanza una precisión del 70,89 % y un área bajo la curva ROC con valor 0,76. En el mismo entorno se encuentran otros algoritmos que tampoco emplean *Information Gain*, y que son el propio *KNN* con K=1 y N-gramas, además de los árboles de decisión con bosques aleatorios empleando Tokens con más de 10 árboles (N=25, N=50, N=75, N=100). Además de éstos, se encuentran *Naïve Bayes*, *SVM* con el núcleo polinomial y *KNN* con K=2.

El resultado más pobre, tanto en precisión como en área bajo la curva ROC es ofrecido por K-vecinos más próximos sin utilizar *Information Gain*, con K=3 y empleando Tokens, que apenas cubre una precisión del 21,25 % y un área bajo la curva ROC igual a 0,24.

5.3.2.6 **Marco primario**

Los resultados de aplicar *Information Gain* en la categorización del marco primario en las conversaciones (véase la Tabla 5.27) arrojan un balance positivo. Todas las técnicas han visto incrementados su precisión y el área bajo la curva ROC, empleando indistintamente N-gramas o tokens.

En términos generales, la familia de algoritmos K-vecinos más próximos ha sido la que más incremento proporcional ha recibido, tanto en precisión como en área bajo la curva ROC. También los árboles de decisión con bosques aleatorios y un número elevado de árboles ha visto mejorada notablemente su precisión. Aun así, no se producen grandes cambios de posición relativa entre los algoritmos una vez aplicada la técnica de IG.

En el conjunto total, el mejor balance del área bajo la curva ROC y de la precisión es el conseguido por el uso de árboles de decisión con bosques aleatorios de 100 árboles ($N = 100$) mediante N-gramas, que obtiene una precisión del 50,36 % y un *AUC* de 0,76. Un resultado similar se consigue a través de la misma técnica pero con 75 árboles (precisión del 49,11 % y AUC de 0,76). Un conjunto de 11 algoritmos logran resultados estadísticamente similares (con precisiones en torno al 50 % y valores del área bajo la curva ROC en el rango de 0,76).

Tabla 5.26: Resultados comparados del uso de *Information Gain* para clasificar la campaña negativa en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	49,64±13,78	0,60±0,15	49,46±12,71	0,59±0,18
	RB: TAN	56,96±11,95	0,62±0,19	50,71±13,37	0,58±0,15
	Naïve Bayes	51,96±12,56	0,50±0,11	36,79±20,05	0,51±0,20
	SVM: Polinomial	50,71± 8,63	0,56±0,10	54,46±13,21	0,65±0,15
	SVM: Polinomial norm.	51,96± 9,39	0,62±0,06	50,71±12,00	0,62±0,09
	SVM: Pearson VII	42,86±16,50	0,53±0,13	48,04±13,87	0,57±0,13
	SVM: RBF	49,46±12,71	0,59±0,12	54,64±15,34	0,62±0,15
	KNN K=1	27,68±12,69	0,51±0,13	29,11±18,63	0,53±0,14
	KNN K=2	26,61± 9,24	0,51±0,10	30,36±15,73	0,52±0,12
	KNN K=3	21,43±10,14	0,52±0,13	30,18±18,64	0,51±0,17
	KNN K=4	22,68±11,33	0,51±0,15	22,68±19,27	0,53±0,19
	KNN K=5	21,43±10,14	0,53±0,11	26,43±13,57	0,52±0,20
	AD: J48	37,14±18,80	0,46±0,18	45,36±22,65	0,59±0,19
	AD: Random Forest N=10	50,36±16,10	0,56±0,12	51,96±11,09	0,53±0,16
	AD: Random Forest N=25	51,96± 9,39	0,56±0,09	49,46±12,71	0,62±0,14
	AD: Random Forest N=50	53,21± 9,92	0,59±0,07	45,71±11,14	0,62±0,11
	AD: Random Forest N=75	50,71±10,45	0,57±0,10	48,21±11,87	0,64±0,11
	AD: Random Forest N=100	51,96± 9,39	0,60±0,11	50,71±10,45	0,63±0,11
Con IG	*RB: K2*	92,50± 8,74	0,86±0,31	88,75±10,94	0,86±0,31
	RB: TAN	87,50±14,43	0,82±0,32	90,00± 9,86	0,82±0,32
	Naïve Bayes	93,75± 6,59	0,78±0,42	92,50±10,54	0,79±0,42
	SVM: Polinomial	91,25± 8,44	0,87±0,32	92,50± 8,74	0,87±0,32
	SVM: Polinomial norm.	93,75± 6,59	0,84±0,32	93,75± 6,59	0,83±0,31
	SVM: Pearson VII	88,75± 3,95	0,89±0,31	90,00± 5,27	0,84±0,31
	SVM: RBF	86,07± 9,21	0,86±0,31	83,57±11,81	0,83±0,31
	KNN K=1	87,32± 8,35	0,74±0,40	86,07± 9,21	0,72±0,40
	KNN K=2	83,57±14,45	0,79±0,37	82,32±16,82	0,77±0,37
	KNN K=3	78,57±17,64	0,83±0,31	77,32±20,15	0,84±0,32
	KNN K=4	77,32±19,27	0,84±0,30	73,57±21,49	0,84±0,31
	KNN K=5	74,82±21,94	0,85±0,31	69,82±22,83	0,85±0,31
	AD: J48	88,75± 9,22	0,71±0,32	88,75± 9,22	0,76±0,33
	AD: Random Forest N=10	92,32± 6,63	0,84±0,33	90,00± 7,91	0,80±0,32
	AD: Random Forest N=25	92,32± 6,63	0,88±0,32	95,00± 6,45	0,80±0,35
	AD: Random Forest N=50	92,32± 6,63	0,87±0,31	93,75± 6,59	0,80±0,35
	AD: Random Forest N=75	93,75± 6,59	0,85±0,31	93,75± 6,59	0,83±0,34
	AD: Random Forest N=100	93,75± 6,59	0,87±0,31	93,75± 6,59	0,84±0,33

Pese a que los *K*-vecinos más próximos ha visto mejorados sus resultados, continúa ofreciendo las precisiones y AUC más pobres del conjunto de resultados cuando no se aplica *Information Gain*.

En definitiva, mediante la aplicación de *Information Gain* se han superado los resultados del modelo básico que no incluía esta técnica y cuyos mejores resultados (Acc = 26,61 % y AUC = 0,57) son muy inferiores a los registrados con la técnica descrita.

5.3.2.7 **Marco secundario**

Los resultados comparados de aplicar *Ganancia de Información* aparecen en la Tabla 5.28. Se aprecia en ellos que el uso de la técnica de *Information Gain* ha permitido mejorar sustancialmente los valores de AUC en la práctica totalidad de algoritmos, así como compensar la gran diferencia que se producía sin *Information Gain* entre la precisión y el área bajo la curva ROC. Con la nueva técnica se avanza positivamente en las ambas direcciones.

Así, la mejor categorización se obtiene utilizando árboles de decisión con bosques aleatorios de 50 árboles mediante el uso de términos y aplicando IG, obteniendo una precisión del 74,82 % y un área bajo la curva ROC de 0,88. Resultados similares (una precisión del 72,32 % y un AUC igual a 0,87) se obtienen con árboles de decisión y *Random Forest* bajo tres posibles condiciones: si se usan N-gramas y 25 ó 100 árboles, o si se usan 75 árboles y usando Tokens.

Partiendo de los resultados más pobres, la familia *K*-vecinos más próximos sufre sustanciales avances porcentuales en precisión cuando se emplean N-gramas, cosa que no sucede con las Tokens. La precisión del resto de algoritmos cuando se utilizan progresa, pero en mucha menor medida, tanto en N-gramas como en Tokens. Los principales avances son en el área bajo la curva ROC, que son más intensos en el caso de empleo de N-gramas.

5.3.2.8 **Análisis del modelo de conversaciones**

Tras analizar los resultados, se puede afirmar que la incorporación de *Information Gain* ha generado una mejora en la respuesta de los algoritmos, tal y como se visualiza en la Tabla 5.30. Así, se han superado de manera notable las carencias de que adolecía el modelo básico de clasificación de los mensajes.

En síntesis, la incorporación de la técnica *Ganancia de Información* ha permitido mejorar sustancialmente la clasificación de la totalidad de las categorías, tanto en precisión como en área bajo la curva ROC. La excepción del marco primario, que apenas mejora en AUC mientras que reduce ligeramente la pre-

Tabla 5.27: Resultados comparados del uso de *Information Gain* para clasificar el marco primario en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	23,93±14,83	0,62±0,17	22,68±14,06	0,63±0,16
	RB: TAN	22,86±14,31	0,61±0,14	21,61±17,85	0,62±0,13
	Naïve Bayes	21,25±14,49	0,56±0,10	17,68±13,37	0,56±0,08
	SVM: Polinomial	15,18± 7,86	0,58±0,07	16,43±14,45	0,61±0,08
	SVM: Polinomial norm.	22,86±10,02	0,63±0,06	21,61±10,50	0,62±0,08
	SVM: Pearson VII	21,61± 8,69	0,58±0,14	24,29±15,59	0,60±0,10
	SVM: RBF	20,36±12,29	0,58±0,11	24,11±12,52	0,63±0,06
	KNN K=1	15,18± 5,21	0,52±0,04	20,18±10,41	0,54±0,08
	KNN K=2	17,68±12,01	0,50±0,05	17,86± 6,99	0,50±0,06
	KNN K=3	16,43±10,23	0,55±0,08	16,61± 8,91	0,49±0,07
	KNN K=4	12,68± 8,35	0,50±0,06	14,11± 9,76	0,47±0,07
	KNN K=5	12,86±10,74	0,50±0,06	14,11±11,40	0,50±0,09
	AD: J48	21,61±10,50	0,54±0,07	25,00±14,43	0,56±0,10
	AD: Random Forest N=10	26,61±14,97	0,57±0,10	23,93±12,26	0,60±0,11
	AD: Random Forest N=25	24,11±11,05	0,63±0,09	20,18±13,34	0,61±0,07
	AD: Random Forest N=50	24,29±16,67	0,60±0,07	18,93±13,42	0,62±0,10
	AD: Random Forest N=75	19,11±10,91	0,61±0,09	21,43±16,62	0,63±0,09
	AD: Random Forest N=100	17,86± 9,14	0,61±0,07	20,18±14,58	0,62±0,11
Con IG	*RB: K2*	42,86±20,27	0,76±0,11	41,61±21,82	0,75±0,10
	RB: TAN	44,29±15,72	0,75±0,09	43,04±14,57	0,75±0,09
	Naïve Bayes	48,04±16,18	0,76±0,09	48,04±12,56	0,75±0,09
	SVM: Polinomial	38,04±20,48	0,71±0,11	39,29±19,97	0,71±0,09
	SVM: Polinomial norm.	39,11±18,79	0,75±0,12	39,11±18,79	0,75±0,10
	SVM: Pearson VII	36,61±20,61	0,78±0,10	36,61±20,61	0,77±0,08
	SVM: RBF	39,11±21,39	0,74±0,10	40,36±19,09	0,75±0,09
	KNN K=1	46,61±17,28	0,71±0,10	46,61±17,28	0,71±0,09
	KNN K=2	39,11±15,78	0,71±0,11	37,86±16,41	0,71±0,09
	KNN K=3	38,04± 8,50	0,70±0,08	35,54±10,15	0,70±0,09
	KNN K=4	41,61± 9,71	0,71±0,09	39,11±12,04	0,72±0,10
	KNN K=5	36,61± 8,80	0,70±0,09	35,36±11,19	0,70±0,10
	AD: J48	35,36±11,19	0,63±0,07	35,71±16,28	0,64±0,11
	AD: Random Forest N=10	40,54±21,08	0,75±0,11	39,11±16,85	0,74±0,08
	AD: Random Forest N=25	47,86±21,51	0,75±0,12	44,11±17,54	0,75±0,08
	AD: Random Forest N=50	45,36±17,46	0,77±0,10	46,61±13,95	0,76±0,09
	AD: Random Forest N=75	49,11±17,62	0,76±0,09	47,86±12,25	0,77±0,09
	AD: Random Forest N=100	50,36±17,14	0,76±0,10	47,86±13,59	0,77±0,09

Tabla 5.28: Resultados comparados del uso de *Information Gain* para clasificar el marco secundario en conversaciones, en términos de precisión(%) y AUC.

		N-gramas		Tokens	
		Prec.(%)	**AUC**	**Prec.(%)**	**AUC**
Sin IG	*RB: K2*	44,46±11,31	0,72±0,11	41,96±12,63	0,72±0,15
	RB: TAN	51,79±11,87	0,75±0,13	49,46±11,26	0,73±0,12
	Naïve Bayes	42,68±18,51	0,67±0,12	50,36±18,13	0,71±0,13
	SVM: Polinomial	49,11±13,10	0,65±0,12	51,96±12,56	0,70±0,14
	SVM: Polinomial norm.	37,86±18,40	0,68±0,14	36,61±14,71	0,73±0,10
	SVM: Pearson VII	23,93±20,69	0,64±0,18	23,93±15,95	0,62±0,17
	SVM: RBF	34,29±16,94	0,57±0,21	37,86±11,41	0,70±0,13
	KNN K=1	12,50±19,54	0,50±0,06	50,18±20,05	0,59±0,08
	KNN K=2	13,75±19,94	0,54±0,07	45,18±23,45	0,57±0,16
	KNN K=3	8,75±19,59	0,55±0,07	32,68±24,28	0,58±0,12
	KNN K=4	8,75±19,59	0,55±0,07	23,93±20,69	0,60±0,13
	KNN K=5	7,50±19,72	0,53±0,09	20,18±22,92	0,60±0,14
	AD: J48	38,21±18,37	0,59±0,14	38,21±18,37	0,57±0,14
	AD: Rand.Forest N=10	41,61±12,80	0,60±0,14	49,82±22,50	0,68±0,16
	AD: Rand.Forest N=25	49,11±15,53	0,67±0,14	50,89±14,37	0,72±0,14
	AD: Rand.Forest N=50	44,29±19,65	0,67±0,17	47,86±13,59	0,78±0,09
	AD: Rand.Forest N=75	49,29±17,82	0,67±0,14	47,86±14,82	0,78±0,10
	AD: Rand.Forest N=100	42,86±19,40	0,67±0,13	46,79±15,40	0,78±0,11
Con IG	*RB: K2*	58,21±14,44	0,85±0,12	59,46±14,19	0,85±0,12
	RB: TAN	59,64±17,17	0,82±0,12	59,64±17,17	0,84±0,11
	Naïve Bayes	56,79±19,27	0,76±0,14	58,21±18,64	0,76±0,13
	SVM: Polinomial	72,14±17,46	0,86±0,11	72,14±17,46	0,86±0,11
	SVM: Polinomial norm.	65,71±16,94	0,87±0,13	66,96±16,09	0,87±0,13
	SVM: Pearson VII	65,54±20,36	0,84±0,13	65,71±17,93	0,84±0,13
	SVM: RBF	62,14±18,40	0,87±0,11	63,39±19,75	0,87±0,12
	KNN K=1	58,21±20,42	0,69±0,15	56,79±20,99	0,69±0,15
	KNN K=2	58,93±21,68	0,70±0,16	58,93±21,68	0,69±0,15
	KNN K=3	41,43±18,38	0,72±0,13	41,43±18,38	0,72±0,12
	KNN K=4	40,18±16,13	0,73±0,12	38,93±15,77	0,72±0,11
	KNN K=5	36,43±20,57	0,75±0,11	36,43±19,71	0,76±0,09
	AD: J48	54,46±11,82	0,68±0,13	50,71±14,61	0,68±0,14
	AD: Rand.Forest N=10	64,46±14,40	0,86±0,13	68,39±15,75	0,85±0,11
	AD: Rand.Forest N=25	72,32±16,28	0,87±0,11	69,64±16,79	0,87±0,12
	AD: Rand.Forest N=50	69,82±15,60	0,86±0,11	74,82±17,55	0,88±0,11
	AD: Rand.Forest N=75	71,07±14,28	0,86±0,11	72,32±17,32	0,87±0,11
	AD: Rand.Forest N=100	72,32±13,99	0,87±0,12	72,32±11,24	0,87±0,11

cisión alcanzada, no impide que el nuevo modelo establece unas condiciones adecuadas para la clasificación en esta categoría.

Tal como se señalaba en el análisis de los resultados de la clasificación de conversaciones en el Capítulo 4, los resultados de precisión y sensitividad obtenidos muestran desviaciones típicas sustanciales. Por esta razón, los resultados adolecen de menor representatividad en al clasificar conversaciones. Es el denominado problema de la variabilidad de los resultados datos por motivos de escasez *data sparsity problem*, que aparece al clasificar textos cortos.

Tabla 5.29: Modelo mejorado de clasificación de conversaciones políticas, en términos de precisión (%) y AUC.

Clasificación	**Algoritmo**	**VSM**	**IG**	**Prec.(%)**	**AUC**
Función informativa	*KNN K=1*	Tokens	Sí	70,89	0,76
Función movilizadora	*RB: K2*	N-gramas	Sí	83,57	0,83
Meta-comunicación	*AD: Random Forest N=100*	Tokens	Sí	70,71	0,83
Campaña negativa	KNN K=1	Tokens	Sí	70,89	0,76
Nivel personalización	*AD: Random Forest N=50*	Tokens	Sí	72,5	0,78
Marco primario	RB:TAN	N-gramas	No	51,79	0,75
Marco secundario	AD: Random Forest N=50	Tokens	Sí	74,82	0,88

Tabla 5.30: Comparativa de mejores resultados de categorización de conversaciones, en términos de precisión (%) y AUC.

	Clasificación	**Algoritmo**	**VSM**	**Prec.(%)**	**AUC**
	Func. informativa	*SVM: Polinom. normalizado*	Tokens	68,57	0,73
	Func. movilizadora	*SVM: Polinomial*	Tokens	71,07	0,79
	Meta-comunicación	*AD: Random Forest N=10*	Tokens	48,21	0,66
Sin IG	*Campaña negativa*	*RB: TAN*	N-gramas	56,96	0,62
	Nivel personalización	*AD: Random Forest N=25s*	N-gramas	*24,11*	*0,63*
	Marco primario	*RB: TAN*	N-gramas	51,79	0,75
	Marco secundario	*SVM: Polinom. normalizado*	N-gramas	*86,25*	*0,6*
	Func. informativa	*KNN K=1*	Tokens	70,89	0,76
	Func. movilizadora	*RB: K2*	N-gramas	83,57	0,83
	Meta-comunicación	*AD: Random Forest N=100*	Tokens	70,71	0,83
Con IG	*Campaña negativa*	*KNN K=1*	Tokens	70,89	0,76
	Nivel personalización	*AD: Random Forest N=50*	Tokens	72,5	0,78
	Marco primario	*AD: Random Forest N=100*	N-gramas	50,36	0,76
	Marco secundario	*AD: Random Forest N=50*	Tokens	74,82	0,88

5.4 **Sumario**

A lo largo de este capítulo se han presentado avances notables en la clasificación automática de mensajes y conversaciones políticas a partir de los resultados empíricos obtenidos en el capítulo 4. Según estos resultados, el modelo construido permite clasificar mensajes y conversaciones políticas que tienen lugar en los sitios sociales de los partidos políticos. En particular, es capaz de reconocer con cierta calidad determinados aspectos de los mensajes y de las conversaciones, con especial relevancia al clasificar las dimensiones funcional, deliberativa, argumentativa y de estrategia comunicativa de los mensajes, aunque esta última con ciertas restricciones. Del mismo modo, el modelo identifica y clasifica conversaciones políticas con relativa calidad en la dimensión funcional, algo menos la dimensión argumentativa y en menor medida la de estrategia comunicativa. No obstante, la capacidad de clasificación del nivel de meta-comunicación, el marcio primario o el marco secundario presentan carencias notables, lo que ha sido objeto de mejora a lo largo del presente capítulo.

En la sección 5.1 se ha presentado la técnica de *Ganancia de Información* o *IG*, que se ha aplicado, en primer lugar, en la mejora de la categorización de los mensajes políticos, descrita en la sección 5.2. En segundo lugar, en la sección 5.3 se ha descrito el empleo de esa misma técnica para ayudar en la mejora de la clasificación de las conversaciones políticas con un doble objetivo. El primero de ellos ha consistido en identificar relaciones entre las clasificaciones etiquetadas características de las propias conversaciones. Para el segundo ha utilizado *Information Gain* para buscar la relevancia de cada una de las características de una conversación que se utilizan a la hora de clasificar cada una de dichas conversaciones, con la consiguiente reducción en tiempo del proceso de clasificación. De este modo se han podido alcanzar resultados notablemente más satisfactorios que los logrados en la primera de las validaciones empíricas.

«El cerebro no es un vaso por llenar,
sino una lámpara por encender»

Plutarco (c. 46 o 50 - c. 120)

CAPÍTULO

6

Conclusiones

A lo largo de este documento hemos ido describiendo la investigación desde su razón de ser hasta la exposición de los resultados alcanzados. Con este capítulo llegamos a la parte final de la tesis, y lo hacemos vinculando el trabajo realizado con la hipótesis fundamental así como con todos los objetivos establecidos a partir de ella, con el fin de evaluar su cumplimiento.

El resto del capítulo está estructurado como sigue. La sección 6.1 resume el conjunto del trabajo realizado. A continuación, la sección 6.2 describe las principales contribuciones realizadas a lo largo de esta tesis, con el resultado final de alcanzar los objetivos de la investigación y validar la hipótesis fundamental. En la sección 6.3 se señalan diversas limitaciones encontradas en el transcurso de la investigación. El trabajo realizado tiene posibles áreas de aplicación que se presentan en la sección 6.4, mientras que en la sección 6.5 se plantean vías de trabajo futuro al hilo del presente trabajo de investigación. Para concluir, el capítulo, la sección 6.6 expone las consideraciones finales de esta tesis doctoral.

6.1 Síntesis del trabajo realizado

Durante esta investigación se ha propuesto un modelo de representación de la acción comunicativa de los partidos políticos en las redes sociales. Para su diseño, se ha llevado a cabo un análisis acerca del estado de la cuestión mediante

una revisión de la literatura en el área. Así, el modelo propuesto se articula sobre cuatro dimensiones que la explican, utilizando los dos elementos centrales de la expresión política: la conversación y el mensaje. A cada uno de ellos se les atribuye distintas dimensiones, que han sido descritas en las secciones anteriores, y que se pueden representar a través de distintas variables.

Por otro lado, se ha creado un modelo que permite clasificar adecuadamente las conversaciones políticas y los mensajes políticos en tanto en cuanto estos representan la acción comunicativa de los partidos. Con este modelo hemos conseguido clasificar con un grado notable de calidad, tanto los mensajes como las conversaciones. Respecto a los mensajes, la calidad a la hora de reconocer el marco primario (dimensión estrategia comunicativa) resulta sensiblemente inferior a la del resto de variables de los mensajes. Por su parte, la categorización de conversaciones presenta una menor capacidad de clasificación, en términos generales, que la correspondiente a los mensajes. Estas dificultades residen principalmente en la identificación del nivel de meta-comunicación (dimensión análisis argumentativo) y de los marcos primario y secundario (dimensión estrategia comunicativa).

A fin de abordar estas dificultades, se ha propuesto una mejora del modelo inicial, por medio de dos vías de mejora. La primera busca identificar posibles relaciones entre las variables que representan el modelo básico de acción comunicativa. Así, incluyendo aquellas que mayor relevancia tienen sobre otras, el modelo ha sido capaz de mejorar su capacidad de reconocimiento. La segunda área ahonda en la necesidad de limitarse en aquella información que representa con mayor nitidez, tanto un mensaje como una conversación. De este modo, aplicando la técnica de *Ganancia de Información* se simplifica la cantidad de información a analizar reduciendo, como consecuencia de ello, el tiempo de respuesta. Con estas adiciones, el modelo consigue obtener unos resultados de mayor calidad, aumentando así su capacidad de reconocimiento.

6.2 Resumen de contribuciones

A lo largo de esta investigación doctoral se han aportado distintas contribuciones, que se enuncian a continuación:

- **Un modelo de representación de la acción comunicativa de los partidos políticos**. El modelo propuesto está basado en las principales dimensiones del comportamiento de los partidos políticos en los sitios web de interacción social, a saber: (i) el nivel funcional, (ii) la dimensión deliberativa democrática, (iii) el nivel argumentativo y (iv) la estrategia de

comunicación. En particular, los mensajes publicados responden a estas cuatro dimensiones, mientras que las conversaciones se representan en tres de ellas: (i) el nivel funcional, (ii) el nivel argumentativo y (iii) la estrategia de comunicación.

- **Un sistema para la recogida de información de los sitios de interacción social**. Mediante este sistema es posible recoger la información de manera automática de los partidos políticos *Partido Popular*, *Partido Socialista Obrero Español* y *Podemos*, pudiendo extenderse para recoger la de otros partidos. Así mismo, es capaz de registrar dicha información de manera interoperable con otros sistemas y facilitar su integración con terceros a través de sistema estándar de gestión de datos tanto en los formatos de ficheros de salida que genera (*ARFF*) para representar conjuntos de datos (de mensajes y de conversaciones) como en el almacenamiento en base de datos relacional.

- **Un modelo de representación de la información de los mensajes publicados en el contexto de conversaciones políticas en los muros digitales de los partidos en la red social *Facebook***. A partir del modelo de representación de la acción comunicativa, de sus dimensiones y de las variables correspondientes a dichas dimensiones se ha elaborado un modelo que permite representar los mensajes en *Facebook*. Este modelo se basa en: (i) características generales del propio mensaje, (ii) características estadísticas extraídas a partir del cuerpo textual del mensaje y (iii) propiedades lingüísticas derivadas de él. Para la obtención del modelo se han utilizado técnicas de recuperación de la información y de procesamiento del lenguaje natural. El modelo de representación resulta necesario para la posterior construcción de un modelo de clasificación de los mensajes.

- **Un modelo de representación de la información de las conversaciones políticas los partidos en redes sociales**. Tal y como sucede con los mensajes, se ha construido un modelo que es capaz de representar la información de dicha acción comunicativa de los partidos. A partir de la información publicada en el sitio social *Facebook*, el modelo representa las principales características de las conversaciones mediante: (i) propiedades generales de la conversación, (ii) propiedades estadísticas de la propia discusión, (iii) propiedades comunicacionales de la conversación, (iv) propiedades del mensaje o *post* inicial y (v) propiedades del flujo de mensajes que se genera a lo largo de la conversación. Este modelo de representación será el que utilice el sistema de clasificación para poder categorizar el conjunto con los datos de las conversaciones.

- **Un modelo básico de clasificación de los mensajes emitidos dentro de las conversaciones políticas**. El modelo construido es capaz de clasificar un mensaje político mediante técnicas de recuperación de información y de aprendizaje automático supervisado a tenor de distintas clasificaciones que responden a las correspondientes dimensiones de su acción comunicativa. Para la dimensión del nivel funcional, las clasificaciones son la función informativa y la función movilizadora. En la dimensión de nivel argumentativo, las clasificaciones son el nivel de personalización, el nivel de meta-comunicación, y el nivel de campaña negativa. Con relación a la dimensión deliberativa democrática, se han determinado las clasificaciones siguientes: el grado de contribución, el grado de debate político y la referencialidad de los contenidos. Finalmente, en la dimensión de estrategia de comunicación, las clasificaciones establecidas son el marco primario y el marco secundario.

- **Un modelo básico de clasificación de las conversaciones políticas que tienen lugar en los muros digitales de los partidos políticos**. Con este modelo, elaborado por medio de técnicas de recuperación de información y de aprendizaje automático supervisado, es posible clasificar una conversación según las categorías que su acción comunicativa establece. En primer lugar, para la dimensión del nivel funcional, las clasificaciones son la función informativa y la función movilizadora. Respecto a la dimensión de nivel argumentativo, las clasificaciones son el nivel de personalización, el nivel de meta-comunicación, y el nivel de campaña negativa. Por último, la dimensión de estrategia de comunicación establece clasificaciones para el marco primario y el marco secundario.

- **Un modelo mejorado de clasificación de la acción automática de las conversaciones políticas y los mensajes vertidos en dichas conversaciones**. Partiendo de las limitaciones que ofrece el modelo básico de clasificación en aspectos como el marco primario de los mensajes, o el nivel de meta-comunicación y los marcos primario y secundario en las conversaciones, se ha diseñado y validado un modelo que permite mejorar la clasificación en estas variables y, consecuentemente, ofrecer una calidad mayor a la hora de clasificar tanto los mensajes en su integridad, como las conversaciones.

Mediante la consecución de estas contribuciones hemos logrado alcanzar los objetivos operacionales enunciados en el Capítulo 1 y que -recordemos- son los siguientes:

Objetivo Operacional 1. *Diseñar y construir un sistema que permite recoger de manera automática la información que representa la comunicación que tiene lugar en los sitios de interacción digital de los partidos.*

Objetivo Operacional 2. *Establecer un modelo de categorización de mensajes publicados en muros políticos.*

Objetivo Operacional 3. *Generar un modelo normalizado que representa la información de los mensajes.*

Objetivo Operacional 4. *Desarrollar y validar un modelo que permite clasificar mensajes en muros digitales de partidos políticos.*

Objetivo Operacional 5. *Establecer un modelo para categorizar las conversaciones políticas en muros políticos.*

Objetivo Operacional 6. *Generar un modelo normalizado que representa la información de las conversaciones.*

Objetivo Operacional 7. *Desarrollar y validar un modelo que permite clasificar conversaciones generadas en los muros digitales de partidos políticos.*

Objetivo Operacional 8. *Definir un método que permite identificar las interdependencias entre las distintas variables que caracterizan a la acción comunicativa.*

Con estos resultados se cumplen los objetivos específicos propuestos al inicio de este documento, como son:

Objetivo específico 1. *Definir un modelo de categorización de la acción comunicativa que ocurre en los sitios de interacción digital de los partidos.*

Objetivo específico 2. *Elaborar un modelo que representa la información de la acción comunicativa en dichos sitios de interacción social.*

Objetivo específico 3. *Diseñar y construir un método que permite validar la categorización de la acción comunicativa generada en los sitios sociales de los partidos políticos.*

Por todo lo anterior, se cumplen el objetivo principal de la tesis:

Objetivo principal 1. *Desarrollar y validar un sistema que describe el comportamiento comunicativo de los partidos políticos en sitios de interacción social online.*

Por todo esto, se puede afirmar que ha sido validad la hipótesis fundamental de la investigación:

> «*Se puede describir el modelo de acción de los partidos políticos en España en sitios de interacción social determinando el comportamiento en su acción comunicativa a través métodos de recuperación de información, minería de datos y técnicas de aprendizaje automático.*»

6.3 Limitaciones

En la presente investigación se plantean algunas restricciones de carácter técnico o de otro orden, que si bien no limitan el alcance de la experimentación ni de las conclusiones, requieren ser explicitados. Estas limitaciones son las siguientes:

- **Política de privacidad de *Facebook***. Pese a que esta red social ofrece transparencia y acceso libre a determinada información que los perfiles pueden definir como pública, su política de privacidad ha venido restringiendo este tipo de acceso. Así, pese a la existencia de una relación de amistad entre dos perfiles, se puede restringir que determinados tipos de informaciones (actualización de datos generales, mensajes, fotos o vídeos) lleguen o no a determinados grupos o incluso personas individuales[1].

[1] Según la política de privacidad de *Facebook*, a la hora de compartir la información de un usuario, por defecto está accesible a las personas con las que éste se comunica o comparte contenido. Se considerará como pública cualquier información que un usuario comparte con el público en general, la información de su perfil público o el contenido que se comparte a través de una página de Facebook o en otro foro público. En todo caso, un usuario puede personalizar el público al que llega la información que publica en su muro o al compartir los contenidos de terceros. Fuente: Política de privacidad de *Facebook* (`https://es-es.facebook.com/about/privacy`)

- **Sesgo de edad**. El principal segmento de edad de personas registradas en *Facebook* está entre los 25 y los 34 años[1]. Pese a que este segmento de población es amplio, no representa al conjunto del sujeto electoral (formado por las personas mayores de edad sin restricciones legales para el ejercicio del voto o sufragio activo), por lo que la representatividad queda limitada.

- **Representatividad de la información de *Podemos*.** El partido político *Podemos* es de reciente creación y por lo tanto, con un bagaje de comunicación inicialmente menor a lo largo del tiempo. Si bien se trata de un partido político no consolidado institucionalmente en el sistema político español, esto ofrece distintas oportunidades [Med15]. Su juventud posibilita nuevas formas de hacer política, sin limitarse al ejercicio de estrategias y prácticas utilizadas por los partidos con más experiencia y gran presencia institucional. En consecuencia, se hace posible el uso de novedosas técnicas de comunicación y acción política en su interacción con la sociedad.

- **Técnicas de aprendizaje supervisado**. Utilizar algoritmos de aprendizaje automático supervisado para el entrenamiento del modelo ofrece riesgos implícitos como el sobreajuste u *overfitting*, según el cual el resultado del aprendizaje puede tener exceso de similitud con respecto a los datos de entrenamiento. Sin embargo, la investigación se centra en el entorno del período electoral, por lo que se asume el posible efecto.

- **Sesgo de partido**. El análisis de la acción comunicativa se ha realizado para el conjunto de tres partidos (*PP*, *PSOE* y *Podemos*) tratado como sistema, y así se han establecido las conclusiones. Ahora bien, la estrategia de cada uno de ellos, su práctica y sus modelos de acción comunicativa pueden ser diferentes, motivo por el cual de los resultados globales no se pueden derivar conclusiones para cada uno de sus miembros, evitando de este modo la *falacia ecológica*.

- **Variabilidad de los modelos de clasificación en procesamiento de textos breves**. Tal como se ha comentado en los análisis de resultados en los capítulos 4 y 5, los resultados de precisión y sensitividad a la hora de clasificar las conversaciones políticas adolecen de una notable dispersión en la clasificación de texto. Recordemos que este fenómeno se presenta a la hora de aplicar técnicas procesamiento mediante clasificación de textos

[1]Fuente: `https://zephoria.com/top-15-valuable-facebook-statistics`

de tamaño reducido, como es el caso de la red social *Twitter*. Es el denominado problema de la variabilidad de los resultados por motivos de escasez de datos o *data sparsity problem*. Como se detalló en el análisis de los resultados, la media [JYZ+11] es el estadístico de referencia que se utiliza en los análisis y comparación de resultados.

6.4 Aplicación a la sociedad

El ámbito en el que esta investigación tiene mayor recorrido es el de los partidos políticos y en particular desde la gestión de la comunicación que desarrollan en relación con los ciudadanos y electores. En este ámbito, puede suponer una ayuda en el desempeño del trabajo de los equipos de comunicación y en especial para el papel de los gestores de comunicación corporativa en redes sociales o *community managers*.

Tanto los modelos de representación de la comunicación, como el sistema para la recogida automática de la información y los correspondientes modelos de clasificación pueden ser la base de futuros desarrollos de aplicaciones informáticas que permitan, tanto obtener una información que sirva para medir el clima político, la detección temprana de temas clave, así como la prevención y la gestión de crisis de comunicación.

Además de los partidos, otros actores políticos como son las instituciones públicas ofrecen un terreno fértil en el que poder aplicar las aportaciones de esta investigación, especialmente en lo relativo al conocimiento de áreas de interés de Gobiernos y Administraciones Públicas. Sus posibles aplicaciones podrían estar orientadas hacia la detección de temas de interés o la valoración *ex-ante* sobre una determinada política pública. En tal caso estaríamos hablando de una herramienta destinada a facilitar el *Gobierno Electrónico* o *e-Government*.

Con un enfoque diferente, a las instituciones o las empresas, en cuestiones relacionadas con la valoración de productos y servicios, o bien para el seguimiento de marca o *brand managing*.

6.5 Trabajo futuro

Al hilo de las contribuciones aportadas y del estado de la cuestión, se plantean diversas líneas por donde continuar la presente investigación. A continuación se describe brevemente cada una de ellas.

6.5.1 **Medir la eficacia real de la comunicación**

Los partidos políticos emplean la comunicación en relación con sus propios objetivos específicos, especialmente de cara al acceso al poder mediante las instituciones políticas o para mantenerse en ellas. Por tanto, lograr el voto es uno de los elementos centrales sobre los que se orienta la comunicación.

En el terreno de los estudios electorales ya se han abordado aspectos como la predisposición a la participación electoral, el propio hecho de votar y el sentido del voto. Cobra, por tanto, interés establecer posibles relaciones entre el desarrollo de la acción comunicativa en los términos que la presente investigación plantea y sus efectos sobre las variables electorales, y en particular el voto. Esta línea de trabajo tiene un carácter más relacionado con la ciencia política que con la aplicación de nuevas técnicas y métodos basados en inteligencia artificial.

6.5.2 **Correspondencia con las redes de difusión e influencia política**

Tal como se describió en el capítulo 2, la influencia de las redes en la sociedad es un elemento central en los sistemas democráticos [Put01]. Se plantean, por ello, dos vías posibles para continuar profundizando la investigación. Por un lado, establecer una posible relación entre las redes políticas creadas a través de sitios web de interacción social y las correspondientes redes sociales presenciales. Por el otro lado, describir la naturaleza, características y efectos de las redes políticas digitales con respecto a la transmisión de información, opiniones o predisposiciones hacia el voto en uno u otro sentido.

En relación a la primera vía, dado el carácter social y político de los redes que se describen, el estudio de los vínculos fuertes o los vínculos débiles que se pueden establecer entre los nodos de una red [BRMA12] se plantea como una vía idónea para progresar en su conocimiento. Además de esto, el trabajo requeriría el uso de otras técnicas de investigación social, tales como las encuestas, bien en su forma presencial o en su modalidad *online*.

Con respecto a la segunda, parece conveniente incorporar el concepto de propagación de la información a través de las redes o en su denominación anglosajona *Information diffusion* [Cho97], que ya ha sido analizado en el ámbito de los blogs [GGLNT04] o en especial las redes sociales digitales, tanto en lo que respecta a la propagación o reproducción de los contenidos como al carácter estable de los mismos [RMK11].

En todo caso, esta línea de investigación futura mantiene el carácter multi-

disciplinar, ahondando en el área de la *social informatics* desde la perspectiva de lo político, la comunicación y el análisis de redes sociales.

6.5.3 Incorporación de meta-información de los contenidos

Los modelos que se han construido para representar la información de la acción comunicativa responden principalmente a la información de carácter textual de las conversaciones, además de aquella otra que se puede convertir en información numérica y que por lo tanto, se puede procesar más fácilmente de manera automática.

Por un lado, la progresiva incorporación de información de carácter semántico permite enriquecer y facilitar la comprensión de la información a nivel cognitivo. Se propone, por ello, incluir información sintetizada (etiquetas o información relevante) obtenida a partir de contenidos no textuales, por ejemplo, vídeo, imágenes, del etiquetado social o *social tagging* [Lam08, NK14, YC15].

6.5.4 Incorporación de detección de sentimientos y representación de entidades

La importancia del factor emocional en la comunicación política tiene recorrido de años [Ric04, Bra06], cuyo interés se ha recuperado en los últimos tiempos [LB08], por lo que supone un área de especial inclinación desde los estudios electorales.

Desde la perspectiva del descubrimiento de conocimiento (*KD*), planteada en el capítulo 2, enriquecer el tratamiento de textos por medio del análisis de sentimientos y minería de opiniones permite analizar las opiniones, actitudes y emociones de las personas hacia las entidades, los individuos, los problemas, los sucesos o temáticas concretas. Tratan de obtener información subjetiva en forma de actitud a partir de un contenido concreto. Esta actitud puede ser un juicio sobre el contenido o bien una emoción. Desde hace años se han abordado estos aspectos subjetivos desde múltiples enfoques. En relación con nuestro área, cabe citar la detección de sentimientos a partir de la expresión en lenguaje informal en redes sociales [TBP+10] o específicamente en el ámbito electoral en la red social *Twitter* [TSSW10].

6.5.5 **Reducción del nivel de variabilidad de los resultados de clasificación**

En relación con una de las limitaciones expresadas, la variabilidad de los modelos de clasificación en procesamiento de textos breves, una de las líneas de trabajo futuro está orientada a la incorporación de técnicas que disminuyan esta variabilidad.

Alguno de los trabajos han abordado esta cuestión, proponiendo distintas alternativas, tales como el uso de técnicas de reemplazo semántico de términos en función del grado de aporte del término al conjunto de la clasificación, como apunta [SHA12], o bien mediante la incorporación de características adicionales del propio cuerpo del mensaje inicial de la conversación. Esta segunda vía está directamente vinculada con la anteriormente mencionada relativa a la mejora mediante de incorporación de meta-información.

6.6 **Consideraciones finales**

Esta investigación busca comprender cómo los partidos actúan comunicando en un nuevo medio como son las redes sociales digitales, donde las personas se expresan cada vez con mayor libertad. Y lo hace poniendo de relieve el papel de los otros actores de la conversación política -los ciudadanos- sobre aspectos como la manera de interpretar las respuestas que ofrecen y los diálogos multilaterales que se establecen como consecuencia de la conversación que inician los primeros.

En un escenario -el digital- siempre cambiante, la tecnología plantea oportunidades y desafíos a actores de los sistemas democráticos como son los partidos y por tanto, sobre el sistema democrático en el que se asientan. Estas demandas se extienden, también hacia las instituciones, en pos de mayores eficacia y transparencia, mediante el uso intensivo de tecnologías de información y comunicación, bien en forma de Administración Electrónica, o bien a los poderes ejecutivos mediante el Gobierno Electrónico.

Simultáneamente, la democracia y la acción política han venido adquiriendo una dimensión en ocasiones performativa, lo que sucede de manera especial durante los procesos electorales. En consonancia con ello, en la comunicación política cobran peso las figuras estilísticas literarias, adquieren notoriedad narrativas colectivas y narrativas coyunturales, las metáforas y otros recursos lingüísticos con el propósito último de comunicar y hacerlo con eficacia al servicio de dichos objetivos electorales.

Internet está llamado a grandes cambios, no ya dentro de años, sino casi de manera inmediata. Innovaciones tecnológicas disruptivas implicarán casi necesariamente cambios en el comportamiento de las personas y en la sociedad, y probablemente también conducirán a transformaciones políticas. En sentido contrario los cambios también tendrán lugar.

La Web es un espacio de normas y usos no explícitos y no cerrados, por lo que el diálogo entre tecnología y sociedad constituye una conversación continua sin solución de continuidad, por lo que en el futuro,

> *«Me preguntarán si soy príncipe o legislador para escribir de política.*
> *Respondo que no y que por eso mismo escribo de política.*
> *Si fuese un príncipe o un legislador*
> *no perdería el tiempo diciendo lo que hay que hacer,*
> *sino que lo haría o me callaría.»*
>
> *Jean-Jacques Rousseau (1712–1778).*
> *El contrato social.*

Bibliografía

[AB02] RÉKA ALBERT AND ALBERT-LÁSZLÓ BARABÁSI. Statistical mechanics of complex networks. *Reviews of modern physics* **74**(1), 47 (2002).

[AC08] NICK ANSTEAD AND ANDREW. CHADWICK. Parties, election campaigning, and the internet toward a comparative institutional approach. *The Routledge handbook of Internet politics* page 56–71 (2008).

[ACC08] EVA ANDUIZA, MARTA CANTIJOCH, AND CAMILO CRISTANCHO. Los ciudadanos y el uso de internet en la campaña electoral. *Las elecciones generales de* (2008).

[ACGS10] EVA ANDUIZA, MARTA CANTIJOCH, AINA GALLEGO, AND JORGE SALCEDO. "Internet y participación política en España". Centro de Investigaciones Sociológicas, Madrid (2010).

[AG05] LADA A. ADAMIC AND NATALIE GLANCE. The political blogosphere and the 2004 us election: divided they blog. In "Proceedings of the 3rd international workshop on Link discovery", pages 36–43. ACM (2005).

[AI95] STEPHEN ANSOLABEHERE AND SHANTO IYENGAR. Going negative: How attack ads shrinks and polarize the electorate (1995).

[AISV94] STEPHEN ANSOLABEHERE, SHANTO IYENGAR, ADAM SIMON, AND NICHOLAS VALENTINO. Does attack advertising demobilize the electorate?. *American political science review* **88**(04), 829–838 (1994).

[Ari07] ARISTÓTELES. Política (2007).

[Arr12] LUIS ARROYO. "El poder político en escena". RBA (2012).

[AW99] S. AMARI AND S. WU. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks* **12**(6), 783–789 (1999).

[Bao00] ANDREW O BAOILL. Slashdot and the public sphere. *first monday* **5**(9) (2000).

[Bar96] GRETCHEN S BARBATSIS. "look, and i will show you something you will want to see": Pictorial engagement in negative political campaign commercials. *Argumentation and Advocacy* **33**(2), 69 (1996).

[Bar05] LARRY M BARTELS. Homer gets a tax cut: Inequality and public policy in the american mind. *Perspectives on Politics* **3**(01), 15–31 (2005).

[BAT91] ROBERT C BERWICK, STEVEN P ABNEY, AND CL TENNY. "Principle-Based Parsing: Computation and Psycholinguistics", volume 44. Springer (1991).

[BCR06] DAVID F BARRERO, J IGNACIO CRIADO, AND M CARMEN RAMILO. Política y web 2.01. (2006).

[BD03] B BIMBER AND R DAVIS. "Campaigning online: The Internet in U.S. elections". Oxford University Press, New York (2003).

[BD11] STANLEY BARAN AND DENNIS DAVIS. "Mass communication theory: Foundations, ferment, and future". Cengage Learning (2011).

[BE07] DANAH BOYD AND NICOLE B. ELLISON. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* **13**(1), 210–230 (2007).

[BE08] D. M BOYD AND NICOLE B ELLISON. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* **13**(1), 210–230 (2008).

[Bec05] ULRICH BECK. "Un nuevo mundo feliz: la precariedad del trabajo en la era de la globalización". Paidós (2005).

[Bem70] DARYL J BEM. "Beliefs, attitudes, and human affairs." Brooks/-Cole (1970).

[Ben95a] LANCE W BENNETT. News: The illusion of politics (1995).

[Ben95b] STEPHEN EARL BENNETT. Comparing americans'political information in 1988 and 1992. *The Journal of Politics* **57**(02), 521–532 (1995).

[Ben03] W LANCE BENNETT. New media power. *Contesting media power* pages 17–37 (2003).

[Ben06a] YOCHAI BENKLER. "The wealth of networks: How social production transforms markets and freedom". Yale University Press (2006).

[Ben06b] SARA BENTIVEGNA. "Campagne elettorali in rete". Laterza (2006).

[Ber86] BERNARD R BERELSON. "Voting: A study of opinion formation in a presidential campaign". University of Chicago Press (1986).

[BF76] MARC BENTON AND P JEAN FRAZIER. The agenda setting function of the mass media at three levels ofinformation holding". *Communication Research* **3**(3), 261–274 (1976).

[BF14] ALBERT-LÀSZLÒ BARABÀSI AND JENNIFER FRANGOS. "Linked: the new science of networks science of networks". Basic Books (2014).

[BFJ+12] ROBERT M BOND, CHRISTOPHER J FARISS, JASON J JONES, ADAM DI KRAMER, CAMERON MARLOW, JAIME E SETTLE, AND JAMES H FOWLER. A 61-million-person experiment in social influence and political mobilization. *Nature* **489**(7415), 295–298 (2012).

[BFP06] ELISA BERTINO, ELENA FERRARI, AND ANDREA PEREGO. Web content filtering. *Web and Information Security* pages 112–132 (2006).

[BGL10] DANAH BOYD, SCOTT GOLDER, AND GILAD LOTAN. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In "System Sciences (HICSS), 2010 43rd Hawaii International Conference on", pages 1–10. IEEE (2010).

[BH00] MICHAEL BERTHOLD AND DAVID I HAND. Intelligent data analysis. *Technometrics* **42**(4), 442–442 (2000).

[BI08] W LANCE BENNETT AND SHANTO IYENGAR. A new era of minimal effects? the changing foundations of political communication. *Journal of Communication* **58**(4), 707–731 (2008).

[Bim98] BRUCE BIMBER. The internet and political mobilization research note on the 1996 election season. *Social Science Computer Review* **16**(4), 391–401 (1998).

[Bim03] BRUCE ALLEN BIMBER. "Information and American democracy: Technology in the evolution of political power". Cambridge University Press (2003).

[Bim08] BRUCE BIMBER. The internet and political fragmentation. *Domestic Perspectives on Contemporary Democracy* page 155 (2008).

[Bio91] FRANK BIOCCA. Some limitations of earlier "symbolic" approaches to political communication. *Television and political advertising* **2**, 11–25 (1991).

[Bis95] CHRISTOPHER M. BISHOP. "Neural Networks for Pattern Recognition". Oxford University Press (1995).

[BK99] J.G. BLUMLER AND D. KAVANAGH. The third age of political communication: Influences and features. *Political Communication* **16**(3), 209–230 (1999).

[BK06] ADAM J BERINSKY AND DONALD R KINDER. Making sense of issues through media frames: Understanding the kosovo crisis. *Journal of Politics* **68**(3), 640–656 (2006).

[BLHL+01] TIM BERNERS-LEE, JAMES HENDLER, ORA LASSILA, ET AL.. The semantic web. *Scientific American* **284**(5), 28–37 (2001).

[Blu51] HERBERT BLUMER. Collective behavior. *New outline of the principles of sociology* pages 166–222 (1951).

[Bor97] PIERRE BORDIEU. Sobre la televisión. *Barcelona: Anagrama* (1997).

[Bra06] TED BRADER. "Campaigning for hearts and minds: How emotional appeals in political ads work". University of Chicago Press (2006).

[BRB02] BERNAD ED BATINIC, ULF-DIETRICH ED REIPS, AND MICHAEL ED BOSNJAK. "Online social sciences." Hogrefe & Huber Publishers (2002).

[Bre01] L. BREIMAN. Random forests. *Machine learning* **45**(1), 5–32 (2001).

[BRMA12] EYTAN BAKSHY, ITAMAR ROSENN, CAMERON MARLOW, AND LADA ADAMIC. The role of social networks in information diffusion. In "Proceedings of the 21st international conference on World Wide Web", pages 519–528. ACM (2012).

[Bru10] JENNIFER BRUNDIDGE. Encountering "difference" in the contemporary public sphere: The contribution of the internet to the heterogeneity of political discussion networks. *Journal of Communication* **60**(4), 680–700 (2010).

[BS02] PAUL R BREWER AND LEE SIGELMAN. Political scientists as color commentators framing and expert commentary in media campaign coverage. *The Harvard International Journal of Press/Politics* **7**(1), 23–35 (2002).

[Bud76] IAN BUDGE. "Party identification and beyond". John Wiley and sons (1976).

[BYRN+99a] RICARDO BAEZA-YATES, BERTHIER RIBEIRO-NETO, ET AL.. "Modern information retrieval", volume 463. ACM press New York. (1999).

[BYRN99b] RICARDO A. BAEZA-YATES AND BERTHIER RIBEIRO-NETO. "Modern Information Retrieval". Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999).

[Cam80] A. CAMPBELL. "The american voter". University of Chicago Press (1980).

[Can98] MARÍA JOSÉ CANEL. Los efectos de las campañas electorales. *Comunicación y sociedad* **11**(1), 47–67 (1998).

[Can99] MARÍA JOSÉ CANEL. Comunicación política. *Técnicas y estrategias para la sociedad de la información. Madrid: Tecnos* (1999).

[Car12] MICHAEL X DELLI CARPINI. Entertainment media and the political engagement. *The SAGE Handbook of Political Communication* page 9 (2012).

[Cas09] MANUEL CASTELLS. "Comunicación y poder". Alianza, Madrid (2009).

[CBHK02] N.V. CHAWLA, K.W. BOWYER, L.O. HALL, AND W.P. KEGELMEYER. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**(3), 321–357 (2002).

[CD85] IVOR CREWE AND DAVID T DENVER. "Electoral change in Western democracies: patterns and sources of electoral volatility". Croom Helm (1985).

[CD07] DENNIS CHONG AND JAMES N DRUCKMAN. Framing theory. *Annu. Rev. Polit. Sci.* **10**, 103–126 (2007).

[CD10] DENNIS CHONG AND JAMES N DRUCKMAN. Dynamic public opinion: Communication effects over time. *American Political Science Review* **104**(04), 663–680 (2010).

[CF96a] STEVEN CHAFFEE AND STACEY FRANK. How americans get political information: Print versus broadcast news. *The Annals of the American Academy of Political and Social Science* pages 48–58 (1996).

[CF96b] ANTHONY CORRADO AND CHARLES M. FIRESTONE. Elections in cyberspace. *Update on L. Related Educ.* **20**, 10 (1996).

[CF10] NICHOLAS A. CHRISTAKIS AND JAMES H. FOWLER. "Conectados". Taurus, Madrid (2010).

[CFME11] M LUZ CONGOSTO, MONTSE FERNÁNDEZ, AND ESTEBAN MORO EGIDO. Twitter y política: Información, opinión y ¿predicción? *Cuadernos de Comunicación Evoca* **4**, 11–15 (2011).

[CGFM12] MICHAEL D CONOVER, BRUNO GONÇALVES, ALESSANDRO FLAMMINI, AND FILIPPO MENCZER. Partisan asymmetries in online political activity. *EPJ Data Science* **1**(1), 1–19 (2012).

[CGH97] E. CASTILLO, J.M. GUTIÉRREZ, AND A.S. HADI. "Expert systems and probabilistic network models". Springer Verlag (1997).

[CGM54] ANGUS CAMPBELL, GERALD GURIN, AND WARREN EDWARD MILLER. "The voter decides". Row, Peterson, and Co. (1954).

[CH91a] G.F. COOPER AND E. HERSKOVITS. A bayesian method for constructing bayesian belief networks from databases. In "Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence", pages 86–94. Morgan Kaufmann Publishers Inc. (1991).

[CH91b] GREGORY F. COOPER AND EDWARD HERSKOVITS. A bayesian method for constructing bayesian belief networks from databases. In "Proceedings of the 1991 conference on Uncertainty in artificial intelligence" (1991).

[Cha06] A. CHADWICK. "Internet politics: States, citizens, and new communication technologies". Oxford University Press, USA (2006).

[Cha07a] ANDREW CHADWICK. Digital network repertoires and organizational hybridity. *Political Communication* **24**(3), 283–301 (2007).

[Cha07b] J. CHAO. Student project collaboration using wikis. In "Software Engineering Education & Training, 2007. CSEET'07. 20th Conference on", pages 255–261. IEEE (2007).

[CHJ03] SHING-LING CHEN, G JON HALL, AND MARK D JOHNS. "Online Social Research: Methods, Issues, & Ethics (Digital Formations, V. 7)". Peter Lang Publishing (2003).

[CHLL09] L CAMAJ, SC HONG, G LANOSGA, AND Y LUO. Political discourse on facebook: A new public sphere. In "annual meeting of the Association for Education in Journalism and Mass Communication, Boston, MA. Retrieved from http://www. allacademic. com/meta/p375074_index. html" (2009).

[Cho97] HUANG CHONGFU. Principle of information diffusion. *Fuzzy Sets and Systems* **91**(1), 69–90 (1997).

[CHY96] MING-SYAN CHEN, JIAWEI HAN, AND PHILIP S. YU. Data mining: an overview from a database perspective. *Knowledge and data Engineering, IEEE Transactions on* **8**(6), 866–883 (1996).

[CILSM09] JAMES CURRAN, SHANTO IYENGAR, ANKER BRINK LUND, AND INKA SALOVAARA-MORING. Media system, public knowledge and democracy a comparative study. *European Journal of Communication* **24**(1), 5–26 (2009).

[CJ96] JOSEPH N CAPPELLA AND KATHLEEN HALL JAMIESON. News frames, political cynicism, and media cynicism. *The Annals of the American Academy of Political and Social Science* pages 71–84 (1996).

[CJ97] JOSEPH N CAPPELLA AND KATHLEEN HALL JAMIESON. "Spiral of cynicism: The press and the public good". Oxford University Press New York (1997).

[CK52] ANGUS CAMPBELL AND ROBERT L KAHN. "The people elect a president". Institute for Social Research University of Michigan (1952).

[CM02] ISMAEL CRESPO MARTÍNEZ. "Las campañas electorales y sus efectos en la decisión del voto: las elecciones generales de 2000 en Españaa". Tirant lo Blanch (2002).

[CMS10] W BRUCE CROFT, DONALD METZLER, AND TREVOR STROHMAN. "Search engines: Information retrieval in practice". Addison-Wesley (2010).

[Col01] STEPHEN COLEMAN. "Elections in the Age of the Internet: Lessons from the United States". Hansard Society London (2001).

[CV95] CORINNA CORTES AND VLADIMIR VAPNIK. Support-vector networks. *Machine learning* **20**(3), 273–297 (1995).

[CYL+08] B.H. CHO, H. YU, J. LEE, Y.J. CHEE, I.Y. KIM, AND S.I. KIM. Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine* **12**(2), 247–256 (2008).

[Dah05a] P. DAHLGREN. The internet, public spheres, and political communication: Dispersion and deliberation. *Political Communication* **22**(2), 147–162 (2005).

[Dah05b] PETER DAHLGREN. The internet, public spheres, and political communication: Dispersion and deliberation. *Political communication* **22**(2), 147–162 (2005).

[Dal84] RUSSELL J DALTON. Cognitive mobilization and partisan dealignment in advanced industrial democracies. *The Journal of Politics* **46**(01), 264–284 (1984).

[DB14] ILARIA DI BONITO. El uso de facebook durante las campañas electorales. reflexiones sobre las elecciones catalanas de 2010 y 2012. *Revista de la Asociación Española de Investigadores en Comunicación* **1**(1) (2014).

[DC08] JOSÉ LUIS DADER AND LIFEN CHENG. Análisis cuantitativo y cualitativo de las webs de los partidos. *Cibercampaña. Cauces y diques para la participación. Las Elecciones Generales de* pages 129–144 (2008).

[DCC+14] JOSÉ LUIS DADER, LIFEN CHENG, EVA CAMPOS, NURIA QUINTANA, AND RICARDO VIZCAÍNO-LAORGA. Las webs de los partidos españoles en campaña electoral. continuismo entre 2008 y 2011. *Trípodos* **1**(34), 115–152 (2014).

[dEPyC] CENTRO DE ESTUDIOS POLÍTICOS Y CONSTITUCIONALES, editor. "Comunidades de partidarios en redes sociales: estudio de las elecciones catalanas de 2010 y 2012".

[DFBA84] RUSSELL J DALTON, SCOTT C FLANAGAN, PAUL ALLEN BECK, AND JAMES E ALT. "Electoral change in advanced industrial democracies: realignment or dealignment?" Princeton University Press Princeton, NJ (1984).

[dG05] E DE GROOTE. "Machine learning in go: Supervised learning in move prediction". Citeseer (2005).

[Dia92] EDWIN DIAMOND. "The spot: The rise of political advertising on television". Mit Press (1992).

[dIS15] CENTRO DE INVESTIGACIONES SOCIOLÓGICAS. Barómetro de enero 2015. Technical Report, Centro de Investigaciones Sociológicas (2015).

[DKWT05] DANIELA V DIMITROVA, LYNDA LEE KAID, ANDREW PAUL WILLIAMS, AND KAYE D TRAMMELL. War on the web the immediate news framing of gulf war ii. *The Harvard International Journal of Press/Politics* **10**(1), 22–44 (2005).

[DL86] RICHARD L DAFT AND ROBERT H LENGEL. Organizational information requirements, media richness and structural design. *Management science* **32**(5), 554–571 (1986).

[dlF10] JORGE RESINA DE LA FUENTE. Ciberpolítica, redes sociales y nuevas movilizaciones en españa: el impacto digital en los procesos de deliberación y participación ciudadana. *Mediaciones sociales* (7), 143–164 (2010).

[DlP14] JORGE DE LA PEÑA. "MACRUSS: Moderación automática de comentarios y reputación de usuarios en sitios web sociales". PhD thesis, Universidad de Deusto (2014).

[DLR77] ARTHUR P DEMPSTER, NAN M LAIRD, AND DONALD B RUBIN. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 1–38 (1977).

[Dow57] ANTHONY DOWNS. An economic theory of political action in a democracy. *The Journal of Political Economy* **65**(2), 135–150 (1957).

[DPD09] DONATELLA DELLA PORTA AND MARIO DIANI. "Social movements: An introduction". John Wiley & Sons (2009).

[DR86] DENNIS K DAVIS AND JOHN P ROBINSON. News story attributes and comprehension. *The main source: Learning from television news* pages 179–210 (1986).

[Dru01] JAMES N DRUCKMAN. The implications of framing effects for citizen competence. *Political Behavior* **23**(3), 225–256 (2001).

[DSG13] PÉNÉLOPE DAIGNAULT, STUART SOROKA, AND THIERRY GIASSON. The perception of political advertising during an election campaign: A measure of cognitive and emotional effects. *Canadian Journal of Communication* **38**(2) (2013).

[DVB06] CLAES H DE VREESE AND HAJO BOOMGAARDEN. News, political knowledge and participation: The differential effects of news media exposure on political knowledge and participation. *Acta Politica* **41**(4), 317–341 (2006).

[dVBS06] CLAES DE VREESE, HAJO BOOMGARDEN, AND HOLLI SEMETKO. Effects of issue news framing on public support for turkish membership in the eu. In "annual meetings of the International Communication Association (ICA), Dresden, Germany", volume 88 (2006).

[Eas53] DAVID EASTON. "The political system". Knopf New York (1953).

[EC93] ALICE H EAGLY AND SHELLY CHAIKEN. "The psychology of attitudes." Harcourt Brace Jovanovich College Publishers (1993).

[Ent07] ROBERT M ENTMAN. Framing bias: Media in the distribution of power. *Journal of communication* **57**(1), 163–173 (2007).

[ESL07] NICOLE B ELLISON, CHARLES STEINFIELD, AND CLIFF LAMPE. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* **12**(4), 1143–1168 (2007).

[Est84] WILLIAM K ESTES. Human learning and memory. In "The biology of learning", pages 617–628. Springer (1984).

[Fag66] RICHARD R FAGEN. "Politics and communication: an analytic study". Little, Brown (1966).

[FD95] RONEN FELDMAN AND IDO DAGAN. Knowledge discovery in textual databases (kdt). In "Proc. 1st Int. Conf. Knowledge Discovery and Data Mining", pages 112–117 (1995).

[Fes62] LEON FESTINGER. "A theory of cognitive dissonance", volume 2. Stanford university press (1962).

[FGG97] N. FRIEDMAN, D. GEIGER, AND M. GOLDSZMIDT. Bayesian network classifiers. *Machine learning* **29**(2), 131–163 (1997).

[FH52] EVELYN FIX AND JOSEPH L HODGES JR. Discriminatory analysis-nonparametric discrimination: Small sample performance. Technical Report, DTIC Document (1952).

[FHR06] LEWIS A FRIEDLAND, THOMAS HOVE, AND HERNANDO ROJAS. The networked public sphere. *Javnost-the public* **13**(4), 5–26 (2006).

[FHT01] JEROME FRIEDMAN, TREVOR HASTIE, AND ROBERT TIBSHIRANI. "The elements of statistical learning", volume 1. Springer Series in Statistics (2001).

[Fis87] DOUGLAS H FISHER. Knowledge acquisition via incremental conceptual clustering. *Machine learning* **2**(2), 139–172 (1987).

[FKM01] DAVID M FARRELL, ROBIN KOLODNY, AND STEPHEN MEDVIC. Parties and campaign professionals in a digital age political consultants in the united states and their counterparts overseas. *The Harvard International Journal of Press/Politics* **6**(4), 11–30 (2001).

[Fos04] SONJA K. FOSS. Rhetorical criticism: Exploration and practice. *Long Grove, Ill* (2004).

[FPSS+96] USAMA FAYYAD, GREGORY PIATETSKY-SHAPIRO, PADHRAIC SMYTH, ET AL.. Knowledge discovery and data mining: Towards a unifying framework. *Knowledge Discovery and Data Mining* pages 82–88 (1996).

[Fre10] DEEN G FREELON. Analyzing online political discussion using three models of democratic communication. *New Media & Society* (2010).

[FSX02] KIRSTEN A FOOT, STEVEN M SCHNEIDER, AND MICHAEL XENOS. Online campaigning in the 2002 us elections. Technical Report, Working Paper (2002).

[Fuc07] CHRISTIAN FUCHS. "Internet and society: Social theory in the information age". Routledge (2007).

[G+95] STEPHEN R GARNER ET AL.. Weka: The waikato environment for knowledge analysis. In "Proceedings of the New Zealand computer science research students conference", pages 57–64. Citeseer (1995).

[Gal03] WILLIAM A GALSTON. If political fragmentation is the problem, is the internet the solution. *The civic web: Online politics and democratic values* pages 35–44 (2003).

[Gaz95] CECILIE GAZIANO. A twenty-five-year review of knowledge gap research. (1995).

[GBBHRM11] SANDRA GONZÁLEZ-BAILÓN, JAVIER BORGE-HOLTHOEFER, ALEJANDRO RIVERO, AND YAMIR MORENO. The dynamics of protest recruitment through an online network. *Scientific reports* **1** (2011).

[GBKB10] S. GONZALEZ-BAILON, A. KALTENBRUNNER, AND R.E. BANCHS. The structure of political discussion networks: a model for the analysis of online deliberation. *Journal of Information Technology* **25**(2), 230–243 (2010).

[GCHS92] WILLIAM A GAMSON, DAVID CROTEAU, WILLIAM HOYNES, AND THEODORE SASSON. Media images and the social construction of reality. *Annual review of sociology* pages 373–393 (1992).

[GD08] JOSÉ LUIS DADER GARCÍA AND ISABEL DÍAZ. Las webs de partidos españoles 2004-2005: una investigación preliminar y de comparación europea, con una propuesta metodológica. In "Estudios de comunicación política: Libro del año 2008", pages 419–444. Asociación Comunicación Política (2008).

[GGLNT04] DANIEL GRUHL, RAMANATHAN GUHA, DAVID LIBEN-NOWELL, AND ANDREW TOMKINS. Information diffusion through blogspace. In "Proceedings of the 13th international conference on World Wide Web", pages 491–501. ACM (2004).

[GGMS80] GEORGE GERBNER, LARRY GROSS, MICHAEL MORGAN, AND NANCY SIGNORIELLI. The "mainstreaming" of america: Violence profile no. 11. *Journal of communication* **30**(3), 10–29 (1980).

[GGP+97] DAN GEIGER, MOISES GOLDSZMIDT, G. PROVAN, P. LANGLEY, AND P. SMYTH. Bayesian network classifiers. In "Machine Learning", pages 131–163 (1997).

[GGR10] JENNIFER GOLBECK, JUSTIN M GRIMES, AND ANTHONY ROGERS. Twitter use by the us congress. *Journal of the American Society for Information Science and Technology* **61**(8), 1612–1621 (2010).

[GH+84] JAMES E GRUNIG, TODD HUNT, ET AL.. "Managing public relations", volume 343. Holt, Rinehart and Winston New York (1984).

[Gid94] ANTHONY GIDDENS. "Consecuencias de la Modernidad". Alianza (1994).

[Git80] TODD GITLIN. "The whole world is watching: Mass media in the making & unmaking of the new left". Univ of California Press (1980).

[GJLS11] BEATRIZ GARCÍA-JIMÉNEZ, AGAPITO LEDEZMA, AND ARACELI SANCHIS. Mmrf for proteome annotation applied to human protein disease prediction. In "Inductive Logic Programming", pages 67–75. Springer (2011).

[GM89] WILLIAM A GAMSON AND ANDRE MODIGLIANI. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology* pages 1–37 (1989).

[GNW03] RACHEL GIBSON, PAUL NIXON, AND STEPHEN WARD. Net gain? political parties and the internet (2003).

[Gof74] ERVING GOFFMAN. "Frame analysis: An essay on the organization of experience." Harvard University Press (1974).

[Gra90] DORIS A GRABER. Seeing is remembering: How visuals contribute to learning from television news. *Journal of communication* **40**(3), 134–156 (1990).

[Gre09] SAMUEL GREENGARD. The first internet president. *Communications of the ACM* **52**(2), 16–18 (2009).

[Gro92] BRUCE E GRONBECK. Negative narratives in 1988 presidential campaign ads. *Quarterly Journal of Speech* **78**(3), 333–346 (1992).

[Gru07] T. GRUBER. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems (IJSWIS)* **3**(1), 1–11 (2007).

[GSGR09] JOSÉ MARÍA GÓMEZ HIDALGO, ENRIQUE PUERTAS SANZ, FRANCISCO CARRERO GARCÍA, AND MANUEL DE BUENAGA RODRÍGUEZ. Web content filtering. *Advances in Computers* **76**, 257–306 (2009).

[Gue08] VASSIA GUEROGUIEVA. Voters, myspace, and youtube. *Social Science Computer Review* **26**, 288–300 (2008).

[Gun87] BARRIE GUNTER. "Poor reception: Misunderstanding and forgetting broadcast news". Routledge (1987).

[GV11] DANIEL GUTIÉRREZ VERA. Ernesto laclau: el populismo y sus avatares. *Iconos. Revista de ciencias sociales* (40), 151–168 (2011).

[GW00] RACHEL GIBSON AND STEPHEN WARD. A proposed methodology for studying the function and effectiveness of party and candidate web sites. *Social Science Computer Review* **18**(3), 301–319 (2000).

[GW12] RACHEL K GIBSON AND STEPHEN WARD. Political organizations and campaigning online. *The SAGE handbook of communication* pages 62–74 (2012).

[Hab89] JÜRGEN HABERMAS. "The Structural Transformation of the Public Sphere". Cambridge (1989).

[Hac84] ROBERT A HACKETT. Decline of a paradigm? bias and objectivity in news media studies. *Critical Studies in Media Communication* **1**(3), 229–259 (1984).

[Hal01] ALEXANDER M CAMPBELL HALAVAIS. "The Slashdot effect: analysis of a large-scale public conversation on the World Wide Web". PhD thesis, (2001).

[Har10] CLAIRE HARDAKER. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research. Language, Behaviour, Culture* **6**(2), 215–242 (2010).

[Hav00] JOHN HAVICK. The impact of the internet on a television-based society. *Technology in Society* **22**(2), 273–287 (2000).

[HB90] CHRISTINA HOLTZ-BACHA. Videomalaise revisited: Media exposure and political alienation in west germany. *European Journal of Communication* **5**(1), 73–85 (1990).

[HCL03] CHIH-WEI HSU, CHIH-CHUNG CHANG, AND CHIH-JEN LIN. A practical guide to support vector classification. (2003).

[HDW94] GEOFFREY HOLMES, ANDREW DONKIN, AND IAN H WITTEN. Weka: A machine learning workbench. In "Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on", pages 357–361. IEEE (1994).

[Hea99] MARTI A HEARST. Untangling text data mining. In "Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics", pages 3–10. Association for Computational Linguistics (1999).

[HH10] JOHN HEILEMANN AND MARK HALPERIN. Game change: Obama and the clintons, mccain and palin, and the race of a. (2010).

[Hin05] MATTHEW HINDMAN. The real lessons of howard dean: Reflections on the first digital campaign. *Perspectives on Politics* **3**(01), 121–128 (2005).

[Hin07] MATTHEW SCOTT HINDMAN. "Voice, Equality and the Internet". Princeton, New Jersey (2007).

[Hin10] MATTHEW HINDMAN. "The myth of digital democracy". Princeton University Press (2010).

[HK06] JIAWEI HAN AND MICHELINE KAMBER. "Data mining: concepts and techniques". Morgan Kaufmann (2006).

[HLL74] JÜRGEN HABERMAS, SARA LENNOX, AND FRANK LENNOX. The public sphere: An encyclopedia article (1964). *New German Critique* pages 49–55 (1974).

[HM87] MARTIN HARROP AND WILLIAM LOCKLEY MILLER. "Elections and voters: a comparative introduction". New Amsterdam Books (1987).

[HM04] DANIEL C HALLIN AND PAOLO MANCINI. "Comparing media systems: Three models of media and politics". Cambridge University Press (2004).

[HMW08] DANIEL HALLIN, PAOLO MANCINI, AND SHEILA WALDECK. "Sistemas mediáticos comparados". Hacer editorial (2008).

[How06] PHILIP N. HOWARD. "New media campaigns and the managed citizen". Cambridge University Press (2006).

[HP05] JACOB S HACKER AND PAUL PIERSON. "Off center: The Republican revolution and the erosion of American democracy". Yale University Press (2005).

[HP09] AUDREY A HAYNES AND BRIAN PITTS. Making an impression: New media in the 2008 presidential nomination campaigns. *PS: Political Science & Politics* **42**(01), 53–58 (2009).

[HS68] HERBERT HIRAM HYMAN AND ELEANOR SINGER. "Readings in reference group theory and research". Free Press New York (1968).

[HSGdBR03] JM GÓMEZ HIDALGO, E PUERTAS SANZ, F CARRERO GARCÍA, AND MANUEL DE BUENAGA RODRÍGUEZ. Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en internet. *Procesamiento del lenguaje natural* **31**, 13–20 (2003).

[Hun05] JEREMY HUNSINGER. Toward a transdisciplinary internet research. *The Information Society* **21**(4), 277–279 (2005).

[HX99] KEI KEUNG HUNG AND LEI XU. Further study of adaptive supervised learning decision (asld) network in stock market. In "IEEE-EURASIP Workshop Nonlinear Signal Image Processing". Citeseer (1999).

[IK10] SHANTO IYENGAR AND DONALD R KINDER. "News that matters: Television and American opinion". University of Chicago Press (2010).

[Ing97] RONALD INGLEHART. "Modernization and postmodernization: cultural, economic, and political change in 43 societies". Princenton University Press (1997).

[Ing15] RONALD INGLEHART. "The silent revolution: Changing values and political styles among Western publics". Princeton University Press (2015).

[Iye87] SHANTO IYENGAR. Television news and citizenséxplanations of national affairs. *American Political Science Review* **81**(03), 815–831 (1987).

[Iye94] SHANTO IYENGAR. "Is anyone responsible?: How television frames political issues". University of Chicago Press (1994).

[Iye10] SHANTO IYENGAR. Experimental designs for political communication research. *Sourcebook for Political Communication Research: Methods, Measures, and Analytical Techniques* pages 129–48 (2010).

[Jac00] WILLIAM G JACOBY. Issue framing and public opinion on government spending. *American Journal of Political Science* pages 750–767 (2000).

[Jac07a] NIGEL JACKSON. Political parties, the internet and the 2005 general election: third time lucky? *Internet Research* **17**(3), 249–271 (2007).

[Jac07b] NIGEL A. JACKSON. Political parties, the internet and the 2005 general election: Third time lucky? *Internet Research* **17**, 249 (2007).

[Jam93] KATHLEEN HALL JAMIESON. "Dirty politics: Deception, distraction, and democracy". Oxford University Press (1993).

[Jen08] HENRY JENKINS. Convergence culture. convergence culture: la cultura de la convergencia de los medios de comunicación/. (2008).

[JL11] NIGEL JACKSON AND DARREN LILLEKER. Microblogging, constituency service and impression management: Uk mps and the use of twitter. *The Journal of Legislative Studies* **17**(1), 86–105 (2011).

[Jon97] KAREN SPÄRCK JONES. "Readings in information retrieval". Morgan Kaufmann (1997).

[JPW01] JUSTIN JACKSON, A PAINTER, AND B WARDLE. E-campaigning: active and interactive. *Painter and Wardle, Viral Politics* pages 142–53 (2001).

[JR07] THORSTEN JOACHIMS AND FILIP RADLINSKI. Search engines that learn from implicit feedback. *Computer* **40**(8), 34–40 (2007).

[JS07] NATALIE JOMINI STROUD. Media effects, selective exposure, and fahrenheit 9/11. *Political Communication* **24**(4), 415–432 (2007).

[JSB00] ARIEL JEREZ, VÍCTOR SAMPEDRO, AND ALEJANDRO BAER. "Medios de comunicación, consumo informativo y actitudes políticas en España". Centro de Investigaciones Sociológicas (2000).

[JYZ+11] LONG JIANG, MO YU, MING ZHOU, XIAOHUA LIU, AND TIEJUN ZHAO. Target-dependent twitter sentiment classification. In "Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1", pages 151–160. Association for Computational Linguistics (2011).

[Kal09] ØYVIND KALNES. Norwegian parties and web 2.0. *Journal of Information Technology & Politics* **6**(3-4), 251–266 (2009).

[Kan08] YAHUI KANG. Knowledge gap effect in health campaign evaluations. In "annual meeting of the International Communications Association, New York, NY" (2008).

[Ken83] JOHN T KENT. Information gain and a general measure of correlation. *Biometrika* **70**(1), 163–173 (1983).

[Kim11] YONGHWAN KIM. The contribution of social network sites to exposure to political difference: The relationships among snss, online political messaging, and exposure to cross-cutting perspectives. *Computers in Human Behavior* **27**(2), 971–977 (2011).

[KK09] MATTHEW J KUSHIN AND KELIN KITCHENER. Getting political on social network sites: Exploring online political discourse on facebook. *First Monday* **14**(11) (2009).

[KL70] ELIHU KATZ AND PAUL FELIX LAZARSFELD. "Personal Influence, The part played by people in the flow of mass communications". Transaction Publishers (1970).

[Kle96] E.M. KLEINBERG. An overtraining-resistant stochastic modeling method for pattern recognition. *The annals of statistics* **24**(6), 2319–2349 (1996).

[Klo04] ROBERT J KLOTZ. "The politics of Internet communication". Rowman & Littlefield (2004).

[Kod99] YVES KODRATOFF. Knowledge discovery in texts: a definition, and applications. *Foundations of Intelligent Systems* pages 16–29 (1999).

[KRS05] ROB KLING, HOWARD ROSENBAUM, AND STEVE SAWYER. "Understanding and communicating social informatics: A framework for studying and teaching the human contexts of information and communication technologies". Information Today, Inc. (2005).

[KS88] JON A KROSNICK AND HOWARD SCHUMAN. Attitude intensity, importance, and certainty and susceptibility to response effects. *Journal of Personality and Social Psychology* **54**(6), 940 (1988).

[KST82] DANIEL KANHEMAN, PAUL SLOVIC, AND AMOS TVERSKY. Judgement under uncertainty: Heuristics and biases (1982).

[L01] DAHLBERG L. The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication, & Society* **4**, 615 (2001).

[LA94] AREND LIJPHART AND DON AITKIN. "Electoral systems and party systems: A study of twenty-seven democracies, 1945-1990". Oxford University Press (1994).

[LA95] AREND LIJPHART AND DON AITKIN. "Sistemas electorales y sistemas de partidos: un estudio de veintisiete democracias, 1945-1990". Centro de Estudios Constitucionales (1995).

[Lak02] GEORGE LAKOFF. "Moral politics: How liberals and conservatives think". University of Chicago Press (2002).

[Lam08] PAUL LAMERE. Social tagging and music information retrieval. *Journal of New Music Research* **37**(2), 101–114 (2008).

[Law04] REGINA G LAWRENCE. Framing obesity the evolution of news discourse on a public health issue. *The Harvard International Journal of Press/Politics* **9**(3), 56–75 (2004).

[LB08] GEORGE LAKOFF AND DAVID BARSAMIAN. "The political mind". Alternative Radio (2008).

[LBG65] PAUL FELIX LAZARSFELD, BERNARD BERELSON, AND HAZEL GAUDET. "The people's choice: how the voter makes up his mind in a presidential campaign". Columbia Univ. Press (1965).

[Lew98] D. LEWIS. Naive (bayes) at forty: The independence assumption in information retrieval. *Machine Learning: ECML-98* pages 4–15 (1998).

[LG11] MIGUEL TÚÑEZ LÓPEZ AND JOSÉ SIXTO GARCÍA. Redes sociales, política y compromiso 2.0: La comunicación de los diputados españoles en facebook. *Revista Latina de comunicación social* (66), 9 (2011).

[LGA+11] GILAD LOTAN, ERHARDT GRAEFF, MIKE ANANNY, DEVIN GAFFNEY, IAN PEARCE, ET AL.. The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication* **5**, 31 (2011).

[LGVG11] GUILLERMO LÓPEZ GARCÍA AND JOSÉ VICENTE GÁMIR. Los medios españoles en la campaña de 2008. In GUILLERMO LÓPEZ GARCÍA, editor, "Política binaria y spam electoral. Elecciones Generales 2008: nuevas herramientasm viejas actitudes". Tirant lo blanch, Valencia (2011).

[LH01] CARL LAGOZE AND JANE HUNTER. The abc ontology and model. In "DC-2001: International Conference on Dublin Core and Metadata Applications 2001", volume 2, pages 1–18. British Computer Society and Oxford University Press (2001).

[Lin87] JUAN JOSÉ LINZ. "La quiebra de las democracias". Alianza Editorial (1987).

[Lin04] JUAN JOSÉ LINZ. Los partidos políticos en las democracias contemporáneas: problemas y paradojas. *Postdata* (10), 187–224 (2004).

[Lip63] SEYMOUR MARTIN LIPSET. "Political man: The social bases of politics", volume 330. Doubleday Garden City, NY (1963).

[LJ10] DARREN G LILLEKER AND NIGEL A JACKSON. Towards a more participatory style of election campaigning: The impact of web 2.0 on the uk 2010 general election. *Policy & Internet* **2**(3), 69–98 (2010).

[LK55] PAUL F LAZARSFELD AND ELIHU KATZ. Personal influence: the part played by people in the flow of mass communications. *Glencoe, Illinois* (1955).

[Lle98] FRANCISCO JOSÉ LLERA. Los rendimientos de los sistemas electorales de las comunidades autónomas: el predominio del bipartidismo imperfecto. *Reis* pages 127–157 (1998).

[Llo82] STUART LLOYD. Least squares quantization in pcm. *Information Theory, IEEE Transactions on* **28**(2), 129–137 (1982).

[LM14] KAY LAWSON AND PETER H MERKL. "When parties fail: emerging alternative organizations". Princeton University Press (2014).

[LR67] SEYMOUR M LIPSET AND STEIN ROKKAN. Cleavage structures, party systems, and voter alignments: an introduction. (1967).

[LR04] CLIFF LAMPE AND PAUL RESNICK. Slash (dot) and burn: distributed moderation in a large online conversation space. In "Proceedings of the SIGCHI conference on Human factors in computing systems", pages 543–550. ACM (2004).

[LS05] ROBERTA LAMB AND STEVE SAWYER. On extending social informatics from a rich legacy of networks and conceptual resources. *Information Technology & People* **18**(1), 9–20 (2005).

[Lue09] ÓSCAR GARCÍA LUENGO. "¿Comunicando desafección?: influencia de los medios en la cultura política". Tecnológico de Monterrey, Escuela de Graduados en Administración Pública y Política Pública (2009).

[Mai02] RANJAN MAITRA. A statistical perspective on data mining. *J. Ind. Soc. Prob. Statist* (2002).

[May01] VICKI MAYER. From segmented to fragmented: Latino media in san antonio, texas. *Journalism & Mass Communication Quarterly* **78**(2), 291–306 (2001).

[MB99] JEFF MANZA AND CLEM BROOKS. "Social Cleavages and Political Change: Voter Alignments and US Party Coalitions: Voter Alignments and US Party Coalitions". Oxford University Press (1999).

[MBM08] S. MAJI, A.C. BERG, AND J. MALIK. Classification using intersection kernel support vector machines is efficient. In "IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", pages 1–8. IEEE (2008).

[McC86] JOHN D MCCARTHY. Prolife and prochoice movement mobilization: Infrastructure deficits and new technologies. *Social Movements and Resource Mobilization in Organizational Society: Collected Essays. New Brunswick, NJ: Transaction Books* (1986).

[Mea76] ROBERT MEADOW. Politics as communication. (1976).

[Med15] JAVIER TORET MEDINA. Una mirada tecnopolítica al primer año de podemos. seis hipótesis. *Teknokultura* **12**(1), 121–135 (2015).

[Mei10] XABIER MEILÁN. "Causas y consecuencias del consumo de información política en España (2000-2009)". Number 68. CIS (2010).

[Mil02] HENRY MILNER. "Civic literacy: How informed citizens make democracy work". UPNE (2002).

[Mil04] KATHERINE MILLER. "Communication theories: Perspectives, processes, and contexts". McGraw-Hill Humanities/Social Sciences/Languages (2004).

[Mit97] TOM M MITCHELL. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill* (1997).

[MM04a] PAUL MCNAMEE AND JAMES MAYFIELD. Character n-gram tokenization for european language text retrieval. *Information Retrieval* **7**(1), 73–97 (2004).

[MM04b] PAUL MCNAMEE AND JAMES MAYFIELD. Jhu/apl experiments in tokenization and non-word translation. *Comparative Evaluation of Multilingual Information Access Systems* pages 85–97 (2004).

[Mor11] DICK MORRIS. "Vote.com: Influence, and the Internet is Giving Power Back to the People". Macmillan (2011).

[MR00] MICHAEL MARGOLIS AND DAVID RESNICK. "Politics as Usual: The CyberspaceRevolution", volume 6. Sage Publications (2000).

[MSLB06] MILLER MCPHERSON, LYNN SMITH-LOVIN, AND MATTHEW E BRASHEARS. Social isolation in america: Changes in core discussion networks over two decades. *American sociological review* **71**(3), 353–375 (2006).

[Mut02a] DIANA C MUTZ. The consequences of cross-cutting networks for political participation. *American Journal of Political Science* pages 838–855 (2002).

[Mut02b] DIANA C MUTZ. Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review* **96**(01), 111–126 (2002).

[Mut06] DIANA C MUTZ. "Hearing the other side: Deliberative versus participatory democracy". Cambridge University Press (2006).

[NBK03] MATTHEW C NISBET, DOMINIQUE BROSSARD, AND ADRIANNE KROEPSCH. Framing science the stem cell controversy in an age of press/politics. *The International Journal of Press/Politics* **8**(2), 36–70 (2003).

[NC] DAN D NIMMO AND JAMES E COMBS. "Mediated political realities".

[Neu92] W RUSSELL NEUMAN. "Common knowledge: News and the construction of political meaning". University of Chicago Press (1992).

[New97] KENNETH NEWTON. Politics and the news media: mobilisation or videomalaise? (1997).

[Nim99] DAN NIMMO. The permanent campaign: marketing as a governing tool. *Handbook of political marketing* pages 73–86 (1999).

[NK14] HYORYUNG NAM AND PK KANNAN. The informational value of social tagging networks. *Journal of Marketing* **78**(4), 21–40 (2014).

[Noh13] DIETER NOHLEN. Controversias sobre sistemas electorales y sus defectos. *Revista española de ciencia política* (31), 9–39 (2013).

[Nor00] PIPPA NORRIS. "A virtuous circle: Political communications in postindustrial societies". Cambridge University Press (2000).

[Nor08] GEOFF NORQUAY. Organizing without an organization: The obama networking revolution. *Options Politiques* page 58 (2008).

[NP99] NORMAN H NIE AND JOHN R PETROCIK. "The changing American voter". iUniverse (1999).

[NWK84] RICHARD G NIEMI, HERBERT F WEISBERG, AND DAVID C KIMBALL. "Controversies in voting behavior", volume 1984. CQ Press Washington, DC (1984).

[Oar97] DOUGLAS W OARD. The state of the art in text filtering. *User modeling and user-adapted interaction* **7**(3), 141–178 (1997).

[O'D01] SUSAN O'DONNELL. Analysing the internet and the public sphere: The case of womenslink. *Javnost-The Public* **8**(1), 39–58 (2001).

[ODG13] LIDIA VALERA ORDAZ, EVA CAMPOS DOMÍNGUEZ, AND GUILLERMO LÓPEZ GARCÍA. ¿ reproduciendo la lógica mediática?: un estudio empírico sobre el impacto del debate electoral español en nuevos y viejos medios. *Perspectivas de la comunicación* **6**(1), 29–49 (2013).

[OM00] NORMAN J ORNSTEIN AND THOMAS E MANN. "The permanent campaign and its future". American Enterprise Institute (2000).

[O'R04] TIM O'REILLY. What is web 2.0. design patterns and business models for the next generation of software. http://oreilly.com/web2/archive/what-is-web-20.html (2004).

[O'R07] T. O'REILLY. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies* (1), 17 (2007).

[Ort11] FÉLIX ORTEGA. "La política mediatizada". Alianza (2011).

[Pap02] ZIZI PAPACHARISSI. The virtual sphere the internet as a public sphere. *New media & society* **4**(1), 9–27 (2002).

[Pat02] THOMAS E PATTERSON. The vanishing voter: Why are the voting booths so empty? *National Civic Review* **91**(4), 367–377 (2002).

[Pat11] THOMAS E PATTERSON. "Out of Order: An incisive and boldly original critique of the news media's domination of Ameri". Vintage (2011).

[Pax99] PAMELA PAXTON. Is social capital declining in the united states? a multiple indicator assessment 1. *American Journal of sociology* **105**(1), 88–127 (1999).

[PD09] JANIS TERUGGI PAGE AND MARGARET E DUFFY. A battle of visions: Dueling images of morality in us political campaign tv ads. *Communication, Culture & Critique* **2**(1), 110–135 (2009).

[Pea82] JUDEA PEARL. "Reverend Bayes on inference engines: A distributed hierarchical approach". Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles (1982).

[PG13] ANDRÉS BOIX PALOP AND GUILLERMO LÓPEZ GARCÍA. Las elecciones europeas de 2009 en españa: el comienzo de la crisis del bipartidismo. *Unión Europea Aranzadi* (11), 25–48 (2013).

[PK97] ZHONGDANG PAN AND GERALD M KOSICKI. Talkshow exposure as an opinion activity. *Political Communication* **14**(3), 371–388 (1997).

[PKV09] NAMSU PARK, KERK F KEE, AND SEBASTIÁN VALENZUELA. Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *CyberPsychology & Behavior* **12**(6), 729–733 (2009).

[PLSSI+12] IKER PASTOR-LÓPEZ, IGOR SANTOS, A SANTAMARIA-IBIRIKA, MIKEL SALAZAR, JORGE DE-LA PEÑA-SORDO, AND PABLO G BRINGAS. Machine-learning-based surface defect detection and categorisation in high-precision foundry. In "Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on", pages 1359–1364. IEEE (2012).

[PM76] THOMAS E PATTERSON AND ROBERT D MCCLURE. "The unseeing eye: The myth of television power in national politics", volume 451. Putnam (1976).

[Poo05] NATHANIEL POOR. Mechanisms of an online public sphere: The website slashdot. *Journal of Computer-Mediated Communication* **10**(2), 00–00 (2005).

[Pos06] NEIL POSTMAN. "Amusing ourselves to death: Public discourse in the age of show business". Penguin (2006).

[PP09] FRANCIS PISANI AND DOMINIQUE PIOTET. "La alquimia de las multitudes: cómo la web está cambiando el mundo". Paidós (2009).

[Pri05] MARKUS PRIOR. News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout. *American Journal of Political Science* **49**(3), 577–592 (2005).

[PSH96] DAVID E PROCTER AND WILLIAM J SCHENCK-HAMLIN. Form and variations in negative political advertising. *Communication Research Reports* **13**(2), 147–156 (1996).

[PT97] VINCENT PRICE AND DAVID TEWKSBURY. News values and public opinion: A theoretical account of media priming and framing. *Progress in communication sciences* pages 173–212 (1997).

[PTP97] VINCENT PRICE, DAVID TEWKSBURY, AND ELIZABETH POWERS. Switching trains of thought the impact of news frames on readers'cognitive responses. *Communication research* **24**(5), 481–506 (1997).

[Put01] ROBERT D. PUTNAM. "Bowling alone: The collapse and revival of American community". Simon and Schuster (2001).

[PZ93] VINCENT PRICE AND JOHN ZALLER. Who gets the news? alternative measures of news reception and their implications for research. *Public opinion quarterly* **57**(2), 133–164 (1993).

[Qui86] J.R. QUINLAN. Induction of decision trees. *Machine learning* **1**(1), 81–106 (1986).

[Qui93] J.R QUINLAN. "C4. 5 programs for machine learning". Morgan Kaufmann Publishers (1993).

[Rae67] DOUGLAS W RAE. "The political consequences of electoral laws". New Haven: Yale University Press (1967).

[Ras97] WAYNE RASH. "Politics on the nets: Wiring the political process". WH Freeman & Co. (1997).

[RBD07] MD M RAHMAN, PRABIR BHATTACHARYA, AND BIPIN C DESAI. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *Information Technology in Biomedicine, IEEE Transactions on* **11**(1), 58–69 (2007).

[RD90] JOHN P ROBINSON AND DENNIS K DAVIS. Television news and the informed public: An information-processing approach. *Journal of Communication* **40**(3), 106–119 (1990).

[Ric04] BARRY RICHARDS. The emotional deficit in political communication. *Political Communication,* **21**(3), 339–352 (2004).

[Ric08] GUILLEM RICO. Monográficas cis 270 (2008).

[RK72] TURNER RALPH AND LEWIS M KILLIAN. Collective behavior. *Englewood Cliffs, NJ* (1972).

[RL04] JOHN D RICHARDSON AND KAREN M LANCENDORFER. Framing affirmative action the influence of race on newspaper editorial responses to the university of michigan cases. *The Harvard International Journal of Press/Politics* **9**(4), 74–94 (2004).

[RM96] PAUL RESNICK AND JAMES MILLER. PICS: Internet access controls without censorship. *Communications of the ACM* **39**(10), 87–93 (1996).

[RM98] RUSSELL D REED AND ROBERT J MARKS. "Neural smithing: supervised learning in feedforward artificial neural networks". Mit Press (1998).

[RMK11] DANIEL M ROMERO, BRENDAN MEEDER, AND JON KLEINBERG. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In "Proceedings of the 20th international conference on World wide web", pages 695–704. ACM (2011).

[RMM+10] CARLOS RUIZ, PERE MASIP MASIP, JOSEP LLUÍS MICÓ, JAVIER DÍAZ-NOCI, AND DAVID DOMINGO. Conversación 2.0. y democracia: análisis de los comentarios de los lectores en la prensa digital catalana. *Comunicación y sociedad= Communication & Society* **23**(2), 7–39 (2010).

[Rob76] MICHAEL J ROBINSON. Public affairs television and the growth of political malaise: The case of "the selling of the pentagon". *American political science review* **70**(02), 409–432 (1976).

[Rok68] MILTON ROKEACH. Beliefs, attitudes and values: A theory of organization and change. (1968).

[RVLET12] JORDI RODRÍGUEZ-VIRGILI, ESTEBAN LÓPEZ-ESCOBAR, AND ANTONIO TOLSÁ. La percepción pública de los políticos, los partidos y la política, y uso de medios de comunicación. (2012).

[RW98] ALEXANDER M ROBERTSON AND PETER WILLETT. Applications ofgrams in textual information systems. *Journal of Documentation* **54**(1), 48–67 (1998).

[Sá01] TERESA SÁDABA. Origen, aplicación y límites de la" teoría del encuadre"(framing) en comunicación. (2001).

[Sam00] VÍCTOR SAMPEDRO. Opinión pública y democracia deliberativa. medios, sondeos y urnas. *International Journal of Media and Cultural Politics* **3**(1) (2000).

[Sam05] VÍCTOR SAMPEDRO. "13M: Multitudes online". Catarata (2005).

[Sam11] VÍCTOR SAMPEDRO. Cibercampaña. cauces y diques para la participación. las elecciones generales de 2008 y su proyección tecnopolítica (2011).

[Sar80] GIOVANNI SARTORI. "Partidos y sistemas de partidos: marco para un análisis". Alianza editorial (1980).

[Sar98] GIOVANNI SARTORI. "Homo Videns: la sociedad teledirigida". Ed. Santillana-Taurus, Madrid (1998).

[Sav02] HEATHER SAVIGNY. Public opinion, political communication and the internet. *Politics* **22**(1), 1–8 (2002).

[Saw05] STEVE SAWYER. Social informatics: Overview, principles and opportunities. *Bulletin of the American Society for Information Science and Technology* **31**(5), 9–12 (2005).

[SBN86] HAROLD W STANLEY, WILLIAM T BIANCO, AND RICHARD G NIEMI. Partisanship and group support over time: A multivariate analysis. *American Political Science Review* **80**(03), 969–976 (1986).

[Sch82] MICHAEL SCHUDSON. "The power of news". Harvard University Press (1982).

[Seb06] FABRIZIO SEBASTIANI. Classification of text, automatic. *The Encyclopedia of Language and Linguistics* **14**, 457–462 (2006).

[Sel98] GARY W SELNOW. "Electronic whistle-stops: The impact of the Internet on American politics". Greenwood Publishing Group (1998).

[SFN08] STEVEN M SMITH, LEANDRE R FABRIGAR, AND MEGHAN E NORRIS. Reflecting on six decades of selective exposure research: Progress, challenges, and opportunities. *Social and Personality Psychology Compass* **2**(1), 464–493 (2008).

[SH95] HERMANN SCHMITT AND SÖREN HOLMBERG. Political parties in decline? (1995).

[Sha51] C.E. SHANNON. Prediction and entropy of printed english. *Bell System Technical Journal* **30**(1), 50–64 (1951).

[Sha91] DONALD LEWIS SHAW. "The rise and fall of American mass media: roles of technology and leadership". Number 2. Roy W. Howard Project, School of Journalism, Indiana University (1991).

[SHA12] HASSAN SAIF, YULAN HE, AND HARITH ALANI. Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org) (2012).

[SHPR00] WILLIAM J SCHENCK-HAMLIN, DAVID E PROCTER, AND DEBORAH J RUMSEY. The influence of negative advertising frames on political cynicism and politician accountability. *Human Communication Research* **26**(1), 53–74 (2000).

[Sim57] HERBERT A SIMON. Models of man; social and rational. (1957).

[Sim08] MARIA SIMONE. Mediated networks for deliberative democracy: Connecting enclave and shared spheres. In "Conference Papers–National Communication Association", pages 1–23 (2008).

[SKM09] YOGESH SINGH, ARVINDER KAUR, AND RUCHIKA MALHOTRA. Comparative analysis of regression and machine learning methods for predicting fault proneness models. *International Journal of Computer Applications in Technology* **35**(2), 183–193 (2009).

[SLTW04] V. SVETNIK, A. LIAW, C. TONG, AND T. WANG. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. *Multiple Classifier Systems* pages 334–343 (2004).

[SM83] G. SALTON AND M.J. MCGILL. "Introduction to modern information retrieval". McGraw-Hill New York (1983).

[SM96] DAVID L SWANSON AND PAOLO MANCINI. "Politics, media, and modern democracy: An international study of innovations in electoral campaigning and their consequences". Greenwood Publishing Group (1996).

[Smi97] STEVE SMITH. Developing new reflexes in framing stories. In "Pew Center/RTNDF workshop "Civic Journalism: Doing it Daily," Chicago. Retrieved February", volume 16, page 2006 (1997).

[Smi09a] AARON SMITH. The internet's role in campaign 2008. *Pew Internet & American Life Project* **15** (2009).

[Smi09b] AARON SMITH. Online participation in the social media era. pew internet & american life project. In "Annual Symposium on Racing & Gaming in Tucson, Arizona. Retrieved on January", volume 20, pages 2011–2032 (2009).

[Smi11] AARON SMITH. The internet and campaign 2010. *Pew Internet & American Life Project* pages 1–39 (2011).

[SN88] HAROLD W STANLEY AND RICHARD G NIEMI. "Vital statistics on American politics". CQ Press (1988).

[SNBN04] DIETRAM A SCHEUFELE, MATTHEW C NISBET, DOMINIQUE BROSSARD, AND ERIK C NISBET. Social structure and citizenship: Examining the impacts of social setting, network heterogeneity, and informational variables on political participation. *Political Communication,* **21**(3), 315–338 (2004).

[Sol] SOLÉ. "Redes complejas: del genoma a Internet".

[SP08] VÍCTOR SAMPEDRO AND FRANCISCO SEOANE PÉREZ. The 2008 spanish general elections:"antagonistic bipolarization" geared by presidential debates, partisanship, and media interests. *The International Journal of Press/Politics* **13**(3), 336–344 (2008).

[SR03] WARD S AND GIBSON R. Online and on message? candidate web sites in the 2001 general election. *The British Journal of Politics and international Relations* **5**, 188 (2003).

[SRJWB86] DAVID A SNOW, E BURKE ROCHFORD JR, STEVEN K WORDEN, AND ROBERT D BENFORD. Frame alignment processes, micromobilization, and movement participation. *American sociological review* pages 464–481 (1986).

[SRVLP08] TERESA SÁDABA, JORDI RODRÍGUEZ-VIRGILI, AND MARÍA TERESA LA-PORTE. La teoría del framing en la investigación en comunicación política. (2008).

[Ste05] LINDA STEINER. The feminist cable collective as public sphere activity. *Journalism* **6**(3), 313–334 (2005).

[Str06] NATALIE TALIA JOMINI STROUD. Selective exposure to partisan information. (2006).

[Sus01] CASS SUSTEIN. "Republic.com". Princeton University Press, Princeton (2001).

[Sus07] CASS SUSTEIN. Republic.com 2.0 (2007).

[SV00] HOLLI A SEMETKO AND PATTI M VALKENBURG. Framing european politics: A content analysis of press and television news. *Journal of communication* **50**(2), 93–109 (2000).

[SW94a] ARIE S SOESILO AND PHILO C WASBURN. Constructing a political spectacle. *The Sociological Quarterly* **35**(2), 367–381 (1994).

[SW94b] W RICHARD STEVENS AND GARY R WRIGHT. "TCP/IP Illustrated: the protocols", volume 1. Addison-Wesley Professional (1994).

[SW06] SOUNDARARAJAN SRINIVASAN AND DELIANG WANG. A supervised learning approach to uncertainty decoding for robust speech recognition. In "Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on", volume 1, pages I–I. IEEE (2006).

[SWY75] GERARD SALTON, ANITA WONG, AND CHUNG-SHU YANG. A vector space model for automatic indexing. *Communications of the ACM* **18**(11), 613–620 (1975).

[Tan01] JAMES W TANKARD. The empirical approach to the study of media framing. *Framing public life: Perspectives on media and our understanding of the social world* pages 95–106 (2001).

[Tar97] SIDNEY TARROW. El poder en movimiento. *Los movimientos sociales, la acción* (1997).

[Tar05] SIDNEY TARROW. "The new transnational activism". Cambridge University Press (2005).

[TB98] STEVEN J TAYLOR AND ROBERT BOGDAN. "Introduction to qualitative research methods: A guidebook and resource ." John Wiley & Sons Inc (1998).

[TBP+10] MIKE THELWALL, KEVAN BUCKLEY, GEORGIOS PALTOGLOU, DI CAI, AND ARVID KAPPAS. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* **61**(12), 2544–2558 (2010).

[Te'01] DOV TE"ENI. Review: A cognitive-affective model of organizational communication for designing it. *MIS quarterly* **25**(2), 251–312 (2001).

[Tel15] FUNDACIÓN TELEFÓNICA. "La Sociedad de la Información en España 2014". Ariel (2015).

[Til78] CHARLES TILLY. "From mobilization to revolution". McGraw-Hill New York (1978).

[TJP+00] DAVID TEWKSBURY, JENNIFER JONES, MATTHEW W PESKE, ASHLEA RAYMOND, AND WILLIAM VIG. The interaction of news and advocate frames: Manipulating audience perceptions of a local public policy issue. *Journalism & Mass Communication Quarterly* **77**(4), 804–829 (2000).

[TK57] RALPH H TURNER AND LEWIS M KILLIAN. Collective behavior. (1957).

[Tor13] JAVIER TORET. Tecnopolítica: la potencia de las multitudes conectadas. el sistema red 15m, un nuevo paradigma de la política distribuida. *IN3 Working Paper Series* (2013).

[TQ12] MARIO TASCÓN AND YOLANDA QUINTANA. "Ciberactivismo: Las nuevas revoluciones de las multitudes conectadas". Los Libros de la Catarata (2012).

[Tri05] JOE TRIPPI. "The revolution will not be televised: Democracy, the Internet, and the overthrow of everything". Regan Books (2005).

[TSFW05] T.C. TURNER, M.A. SMITH, D. FISHER, AND H.T. WELSER. Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication* **10**(4), 00–00 (2005).

[TSSW10] ANDRANIK TUMASJAN, TIMM O SPRENGER, PHILIPP G SANDNER, AND ISABELL M WELPE. Predicting elections with twitter: What 140 characters reveal about political sentiment. In "Proceedings

of the fourth international aaai conference on weblogs and social media", pages 178–185 (2010).

[TWPL06] KAYE D TRAMMELL, ANDREW PAUL WILLIAMS, MONICA POSTELNICU, AND KRISTEN D LANDREVILLE. Evolution of online campaigning: Increasing interactivity in candidate web sites and blogs through text and technical features. *Mass Communication & Society* **9**(1), 21–44 (2006).

[ÜMB07] B. ÜSTÜN, WJ MELSSEN, AND LMC BUYDENS. Visualisation and interpretation of support vector regression models. *Analytica chimica acta* **595**(1-2), 299–309 (2007).

[VB11] JM VALLÉS AND X BALLART. Política para apolíticos. *Barcelona, Editorial SA* (2011).

[VBVAdV10] RENS VLIEGENTHART, HAJO G BOOMGAARDEN, PETER VAN AELST, AND CLAES H DE VREESE. Covering the us presidential election in western europe: A cross-national comparison. *Acta Politica* **45**(4), 444–467 (2010).

[Ver05] MIEKE VERLOO. Mainstreaming gender equality in europe. a critical frame analysis approach. *The Greek Review of Social Research* **117**(11-34) (2005).

[VJ12] MANUEL VILLORIA AND FERNANDO JIMÉNEZ. La corrupción en españa (2004-2010): datos, percepción y efectos/corruption in spain (2004-2010): Data, perception and consequences. *Reis* pages 109–134 (2012).

[VO12] LIDIA VALERA ORDAZ. La campaña en la red: un estudio de la agenda temática y los marcos de pp y psoe en las elecciones de 2011. In "Actas del IV Congreso Latina de Comunicación Social. Universidad de la Laguna", volume 4, (2012).

[VSDV99] PATTI M VALKENBURG, HOLLI A SEMETKO, AND CLAES H DE VREESE. The effects of news frames on readers'thoughts and recall. *Communication research* **26**(5), 550–569 (1999).

[WC00] VINCENT WAN AND WILLIAM M CAMPBELL. Support vector machines for speaker verification and identification. In "Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop", volume 2, pages 775–784. IEEE (2000).

[Web98] JANE WEBSTER. Desktop videoconferencing: experiences of complete users, wary users, and non-users. *MIS quarterly* pages 257–286 (1998).

[Wes97] DARRELL M WEST. "Air wars: Television advertising in election campaigns, 1952-1996". Cq Press (1997).

[Wes07a] MIKE WESTLING. Expanding the public sphere: The impact of facebook on political communication. *Society* **28**(May), 835–860 (2007).

[Wes07b] MIKE WESTLING. Expanding the public sphere: The impact of facebook on political communication. *The New Vernacular* (2007).

[Wes13] DARRELL M WEST. "Air Wars: Television Advertising and Social Media in Election Campaigns, 1952-2012". Sage (2013).

[WG09] CHRISTINE B WILLIAMS AND GIRSH J GULATI. Facebook grows up: An empirical assessment of its role in the 2008 congressional elections. *Proceedings from Midwest Political Science Association, Chicago* (2009).

[WGC+98] SUSAN WARD, RACHEL K GIBSON, I CREWE, B GOSSCHALK, AND J BARTLE. The first internet revolution? united kingdom political parties and campaigning in cyberspace. In "Political Communications: Why Labour Won the General Election of 1997". Frank Cass, Londres (1998).

[Wil70] ROBIN MURPHY WILLIAMS. "American society: A sociological interpretation". Knopf New York (1970).

[Wil97] YORICK WILKS. Information extraction as a core language technology. *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology* pages 1–9 (1997).

[WM09] MAGDALENA E WOJCIESZAK AND DIANA C MUTZ. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of Communication* **59**(1), 40–56 (2009).

[WPF+09] ANDREW SCOTT WAUGH, LIUYI PEI, JAMES H FOWLER, PETER J MUCHA, AND MASON ALEXANDER PORTER. Party polarization in congress: A network science approach. (2009).

[WS92] W.J. WILBUR AND K. SIROTKIN. The automatic identification of stop words. *Journal of information science* **18**(1), 45–55 (1992).

[YC15] CHIN-SHENG YANG AND PEI-CHUN CHANG. Mining social media for enhancing personalized document clustering. In "HCI in Business", pages 185–196. Springer (2015).

[YP98] SIMEON J YATES AND JANE L PERRONE. Politics on the web. *ASSIGNATION-LONDON-* **15**, 49–53 (1998).

[Zal92] JOHN ZALLER. "The nature and origins of mass opinion". Cambridge university press (1992).

[Zij79] ANTON C ZIJDERVELD. "On cliches: The supersedure of meaning by function in modernity". Routledge and Kegan Paul London (1979).

[Zuc05] ALAN S ZUCKERMAN. "The social logic of politics: Personal networks as contexts for political behavior". Temple University Press (2005).

Esta tesis doctoral se terminó de escribir en
Bilbao el 27 de agosto de 2015.