From Forensic to Preventive: Language Agnostic Approach to Preventive Detection of Malicious Users

PhD dissertation by Rubén Sánchez Corcuera

within the doctoral program Engineering for the Information Society and Sustainable Development

Candidate	Advisors	
Rubén Sánchez Corcuera	Dr. Aitor Almeida Escondrillas	Dr. Arkaitz Zubiaga Mendialdua

Para ti, abuela.
Porque tu sonreirías al verla terminada.

Abstract

Malicious users have been active on Twitter for over a decade, using various strategies to make their attacks work. However, it was not until 2016 when it was detected that this type of user could break into electoral processes, such as the US presidential election of that year or that they could create attack vectors against users or groups through the exploitation of public information. Following the detection of these events, researchers in the area and the social network itself focused on creating better models capable of detecting these users, whom they called social or sybil bots.

However, the models presented so far have not been able to put an end to malicious users due to their constant evolution, the approach used and the specialisation in automated users or bots. The approach used so far, called forensics, makes use of data from completed attacks to train detection models. Therefore, when used on datasets where attacks have not been carried out, the models are not able to detect potentially malicious users. Moreover, most of the systems developed by the scientific community have focused on detecting or characterising bots, leaving out other users, managed in other ways, who also take malicious actions against legitimate ones.

For this reason, at the beginning of this doctoral thesis, we have conducted two analyses in which we demonstrate the ease with which attack vectors can be generated by exploiting public information on Twitter. Subsequently, we present the development and evaluation of a model capable of exploiting user interactions to

foresee their future behaviour and classify malicious users, regardless of how they are managed. The system presents a preventive approach that aims to detect malicious users before they carry out their attacks. To this end, the model is trained and evaluated using official attack data published in the Twitter Transparency Centre and respecting the timeliness of the data so that it can be used in real situations. Finally, the implemented system is agnostic to the language and biographical characteristics of the users so that it can be used on data from different countries and is not dedicated to a specific type of user.

Resumen

Los usuarios maliciosos llevan actuando en Twitter desde hace más de una década utilizando diversas estrategias para que sus ataques funcionen. Sin embargo, no fue hasta 2016 cuando se detectó que este tipo de usuarios eran capaces de irrumpir en procesos electorales como en la elección a la presidencia de Estados Unidos de dicho año o que eran capaces de crear vectores de ataque contra usuarios o colectivos a través de la explotación de información pública. Tras la detección de estos sucesos los investigadores en el área y la propia red social centraron sus esfuerzos en crear mejores modelos capaces de detectar a estos usuarios a los que denominaron social bots o sybil bots.

Sin embargo, los modelos presentados hasta el momento no han sido capaces de terminar con los usuarios maliciosos debido la constante evolución de estos, al enfoque utilizado y la especialización en usuarios automatizados o bots. El enfoque utilizado hasta el momento, denominado forense, hace uso de datos de ataques finalizados para entrenar los modelos de detección. Por ello, al ser utilizados en conjuntos de datos en los que los ataques no han sido llevados a cabo, los modelos no son capaces de detectar a los usuarios potencialmente maliciosos con antelación. Por otra parte, la mayoría de los sistemas desarrollados por la comunidad científica se han centrado en la detección o caracterización de bots por lo que dejan de lado a otros usuarios, gestionados de otra manera, que también emprenden acciones maliciosas contra los legítimos.

Es por ello por lo que al comienzo de tesis doctoral hemos llevado a cabo dos análisis en los que se demuestra la facilidad con la que se pueden generar vectores de ataque mediante la explotación de información pública en Twitter. Posteriormente se presenta el desarrollo y evaluación de un modelo capaz de explotar las interacciones entre usuarios para prever el comportamiento que estos tendrán en el futuro y clasificar así a los usuarios maliciosos, independientemente de cómo estén gestionados. El sistema presenta un enfoque preventivo con el que se pretende detectar a los usuarios maliciosos antes de que estos lleven a cabo su ataque. Para ello, el modelo se entrena y evalúa utilizando datos de ataques oficiales publicados en el Centro de Transparencia de Twitter y respetando la temporalidad de los datos para que pueda ser empleado en situaciones reales. Finalmente, el sistema implementado es agnóstico del lenguaje y de las características biográficas de los usuarios por lo que puede ser empleado sobre datos de diferentes países y no está dedicado a un tipo especifico de usuarios.

Acknowledgements

Si he llegado a escribir estas líneas ha sido por el apoyo de muchas personas que han pasado por mi vida durante estos tres años y me han moldeado hasta convertirme en la persona que soy hoy.

Lo primero de todo me gustaría agradecer a mi familia por el apoyo incondicional que he recibido siempre. A ama y aita por confiar en mí, por dejarme tomar mis propias decisiones, por estar ahí cuando me he equivocado y por decirme siempre lo que piensan que es mejor para mí. A Carla por ser mi cómplice y compañera en casa durante muchos años y abrirme camino en muchos aspectos. A ti, Amparo, a quien va dedicado este logro, por ser madre, abuela, cocinera, costurera, profesora y unión de una familia. Porque, aunque te fuiste sin verla terminada, sé que estarías orgullosa de saber que tu nieto ha terminado, por fin, esta aventura. Y a toda la familia que ha estado a mi lado durante este tiempo y a la que he tenido que explicar mil veces de que iba todo este tema.

Por otra parte, me gustaría agradecer a mis directores Aitor Almeida y Arkaitz Zubiaga por todo el apoyo que me han ofrecido a lo largo de estos años. Por ayudarme cuando he caido en el "pozo" y ofrecerme soluciones a los atascos encontrados en el camino. También a todas las personas de MORElab por acogerme con tanto cariño cuando llegué y por ayudarme con todo lo que he necesitado a lo largo de estos años. En especial a Zulaika y Aritz por ser un apoyo firme desde el inicio y compañeros en todas las tonterias que se me han ocurrido durante estos años y a Oihane por poner un poco de cabeza a este trio.

También quiero acordarme de la cuadrilla de toda la vida: Asier, Diego, Gaizka, Iker y Xabi. Gente importante en mi vida, dispuestos a tener conversaciones difíciles y a ayudarme cuando lo he necesitado. ¡Por todo lo que habéis hecho por mí, esta tesis también es por vosotros! Tampoco quiero olvidarme de mis queridos Olivenses, por sacarme de Bilbao y de mi zona de confort para coger aire y volver renovado. Ni tampoco de toda la gente de Itaka que ha estado cerca de mi durante mi camino y que ha confiado en que aun siendo un "pieza" llegaría a terminar este documento.

Finalmente, quiero dar las gracias a Gabriela, por llegar como un terremoto a removerme la vida, como un torrente de sinceridad, cariño y felicidad que me ha estado empujando durante toda la escritura de esta tesis.

Termino igual que he empezado, dando las gracias a toda la gente que ha pasado por mi vida durante estos años. Se que me dejo a muchas personas que han hecho estos años tal y como han sido, a todas ellas...

Eskerrik asko
Rubén Sánchez Corcuera

Table of Contents

List of Figures

List of Tables

Acronyms

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

CNN Convolution Neural Networks

DMCA Digital Millennium Copyright Act

DTC Decision Tree Classifier

GAN Generative Adversarial Networks

GAT Graph Attention Network

GCN Graph Convolutional Network

GNN Graph Neural Network

HITS Hyperlink-Induced Topic Search

KNN Nearest Neighbours

LSTM Long Short-Term Memory

MLP Multi-Layer Perceptron

NB Naive Bayes

NGO Non-Governmental Organizations

NLP Natural Language Processing

OSN Online Social Network

OSoMe Observatory on Social Media

PRC People's Republic of China

ReLU Rectified Linear Unit

RF Random Forest

RNN Recurrent Neural Network

RQ Research Question

SVC Support Vector Classifier

TGN Temporal Graph Network

TMRC Twitter Moderation Research Consortium

TPL Tweets Processing Library

TTC Twitter Transparency Center

URL Uniform Resource Locator

WandB Weight and Biases

"Once an idea has taken hold of the brain it's almost impossible to eradicate".

Cobb - Inception

CHAPTER

1

Introduction

Online Social Networks (OSNs) were created to connect people regardless of their location or intention. These platforms have revolutionised how people interact, find jobs or get information and have even changed how organisations communicate their latest news to their clients or members. More than a decade after their conception, the main OSNs are still growing, and many more emerged as they continue to prove that they are one of the best tools for delivering content of all kinds to people of all ages (Kapoor et al., 2018; Aichner et al., 2021). However, due to their popularity and the thousands of daily users, they are perfect places to be used as platforms for attacks conducted by malicious users or trolls that aim to harm legit users. Several strategies can be followed to achieve the objective behind the attacks; fake news spamming (Zhang and Ghorbani, 2020), misinformation (Suarez-Lledo et al., 2021), or astroturfing (Ratkiewicz et al., 2011) are some famous examples on Twitter. Even though the attacks are being carried out on many social networks, the microblogging social network Twitter is the platform on which researchers have focused to carry out studies on the attacks (Pacheco et al., 2020b,a; Zannettou et al., 2019), and also the social network itself has created several reports about them (Twitter, 2020). Twitter, created in 2006, is one of the world's most

popular social networking sites. Twitter is a microblogging social network created to provide a platform for short texts of up to 280 characters, called tweets.

These attacks, even if they are being carried out in an online environment, have managed to alter processes of utmost importance, such as the extensively analysed 2016 USA presidential elections (Bessi and Ferrara, 2016; Bovet and Makse, 2019). In it, thousands of coordinated accounts were detected to change US citizens' voting intentions using idea induction techniques. In Europe, induction attacks are common; for example, many accounts dedicated to changing the vote intention were detected during the Brexit campaign, also held in 2016 (Bastos and Mercea, 2019; Mora-Cantallops et al., 2019). In Spain, an analysis of tweets generated from more than 750.000 users in the context of the national elections of 2019 demonstrated the presence of more than 40.000 social bots supporting all of the five main political parties by generating a high volume of daily interactions (Pastor-Galindo et al., 2020). These are examples of politically driven attacks with an impact on society. However, attacks are also perpetrated without political intentions, such as coordinated anti-vaccine movements (Broniatowski et al., 2018; Yuan et al., 2019) or disinformation attacks on COVID-19 (Rosenberg et al., 2020; Kouzy et al., 2020).

However, although the models proposed by researchers in the area report excellent results in detecting bots on Twitter, all approaches to date are focused on detecting the so-called social bots or cyborgs, leaving aside other malicious users such as regular human accounts. Social bots or cyborgs are characterised as accounts fully or partially controlled by algorithms that may mimic human behaviour to deceive legit users. Focusing only on these accounts may leave behind malicious accounts that carry attacks on social media, as proven by (Bevensee and Ross, 2018). Furthermore, all approaches of bot detection to date are forensic, i.e. they train their models using datasets of finished attacks. Thus, the models learn to identify profiles that have already finished the attacks. In this PhD dissertation, we try to address the early detection of malicious users, which we define as Twitter users who try to conduct attacks against other users or ideas regardless of how they are managed, i.e.,

bots, cyborgs, or humans. To this end, we envisage the creation of preventive systems that can alert users to malicious users or the danger of an attack before it is carried out. This is proposed by implementing a model capable of predicting the actions of users on Twitter and trained to profile whether users are malicious or not. Detecting or identifying attacks or malicious users as quickly as possible can help create freer, fairer and more objective social networks where people can express themselves without being biased in their opinions.

The remainder of this chapter is structured as follows: Section 1.1 explains the context and motivation behind the research. Afterwards, Section 1.2 formulates the hypothesis, objectives and scope of the work. Section 1.3 describes the followed research methodology to achieve the objectives, and Section 1.4 summarises the principal scientific and technical contributions of this dissertation. The chapter concludes with an outline of the dissertation in Section 1.5.

1.1 Context and Motivation

The problems and, thus, consequences these attacks bring to social network users and their impact in important events in societies has encouraged a prolific research area in this topic. Even though there are several research fields related to the security in social networks (fake news, rumours, bot detection, etc.), in this thesis, we focus on the detection, identification and characterisation of the so called malicious users.

Researchers in the area of malicious user detection have dedicated their efforts to analyse the attacks and the users who cause them. The result of these investigations has been the profiling and definition of the attackers, who have been given the name of *social bots* by the scientific community. This name is because many of these users are artificial profiles controlled by computers using algorithms that mimic the behaviour of humans in order to create a feeling of closeness to legit users. Social bots are created as fake humans to take advantage of the principle of homophily. As stated in (McPherson et al., 2001) *"Homophily is the principle that states that a contact between*

similar people occurs at a higher rate than among dissimilar people". Using accounts similar to the ones they want to attack and leveraging the fact that emotions are contagious on social networks (Kramer et al., 2014), attackers get to mislead users and get their message more compelling.

Due to the use of such strategies that allow malicious users to carry out their attacks successfully, the related scientific community started to develop automated user detection systems. However, the first detection systems did not differentiate between so-called social bots with malicious intentions and other automated users, such as accounts that allow downloading videos uploaded to Twitter (@this_vid). To avoid including legitimate bots as social bots, and as social bots evolved to better camouflage themselves among legitimate users, researchers developed new systems to detect these users using features that would better identify them, such as relationships with other users or the speech and sentiment they were using in tweets. The most important work on detecting bots on Twitter, called Botometer, is presented by researchers at Observatory on Social Media (OSoMe), Indiana University (Sayyadiharikandeh et al., 2020). This software was presented in the competition that DARPA held in 2016 to find the best software to identify bots on Twitter. This competition was held because of the increasing attacks against the government and the presence of terrorist groups flooding the social web with camouflaged propaganda (Subrahmanian et al., 2016). The proposed model by researchers at OSoMe makes use of thousands of features to classify users between bots and humans. These features are extracted from the account's profile, friends, social network structure, temporal activity patterns, and tweets and processed and introduced in several ensembles of machine learning models to compute the bot score for the selected account. Despite Botometer being the most relevant model, the scientific community has proposed many solutions that using a different set of features and algorithms achieves very good results in the classification (Lingam et al., 2019; Stieglitz et al., 2017; Wu et al., 2020; Cresci et al., 2019b).

However, automated users with malicious intent continue to act on Twitter and carry out these attacks undetected. This is due to the approach that the scientific community uses to detect social bots. Following the division of

approaches made by (Milon-Flores and Cordeiro, 2022), there are two types of approaches depending how they use the data to train their models. The first approach is called forensic in which the data is used without respecting temporality, i.e. data from the beginning to the end of the attack are used for training the model. In contrast, in the preventive model, the temporality of the data is respected when training. In this approach, a point in the dataset is set to represent the time at which we want to detect malicious users and only the data prior to that time can be used to train our models. In this way, the models learn the previous steps performed by malicious users and are able to detect them before they carry out the attack.

Therefore, for the beginning of this thesis, we have decided to conduct two initial experiments to demonstrate the ease with which inference of private information can generate attack vectors for users and organisations on Twitter. The main objective of these two experiments is to simply demonstrate the vulnerability that users experience on this social network by interacting with other users and posting information about them. Subsequently, and after demonstrating the risk of exposure, we have developed a system with a preventive approach capable of predicting the actions that Twitter users will take in the future and subsequently classifying their representations to detect malicious ones.

This thesis aims to devise a set of best practices to inform the design of social networking platforms that are less prone to be manipulated and controlled by corporations or third parties capable of creating these attacks. In this way, social networks can be built in which users can inform themselves or give their opinion more unrestrainedly and without being conditioned by the bombardment of information that these attacks carried out. We also believe that creating these free and fair social networks directly influences societies as citizens do not suffer from the induction of ideas that can alter their opinions.

1.2 **Hypothesis, Objectives and Scope**

Based on the current state of malicious user detection on Twitter, the hypothesis of this dissertation is:

Hypothesis. *Using interactions that users make on Twitter, it is possible to create a language agnostic system that foresees malicious users' behaviour and detects them in a preventive way.*

To be able to validate this hypothesis, the general goal of this dissertation is:

Goal. *Design, implement and validate a language agnostic model capable of foreseeing users' behaviour on Twitter to detect possible malicious users who intend to deploy idea induction attacks.*

This general goal can be achieved by addressing the following more specific and measurable objectives:

O1 To study the current state of the art on malicious user profiling on OSNs and the automated malicious user detection systems on Twitter.

O2 To analyse and leverage the open source intelligence on Twitter to infer information from communities on Twitter.

- **O2.a** To design, implement and evaluate a model for inferring the hierarchical roles of users in the organisations from which they are members using information from their Twitter relationships.
- **O2.b** To design, implement and evaluate a model capable of differentiating organisation-specific languages using Twitter data from their members.

O3 To design and implement a time-unrestricted Twitter data capture tool to complement the data provided by the Twitter Transparency Center (TTC).

O4 To identify an appropriate evaluation methodology for the malicious user foreseeing model task with its correspondent metrics and perform a quantitative analysis of the results.

O5 To design, implement and evaluate a supervised deep learning model that employs interactions between Twitter users to foresee their behaviour and classify them as malicious users pre-emptively.

The resulting malicious user foreseeing system should also fulfil the following requirements:

1. *Language agnostic:* as the malicious user foreseeing model does not employ the tweet's content to make its classification, the model stays language agnostic and can be used with data from any location.

2. *Interaction based:* the malicious user foreseeing model is based only on interactions between users and therefore characterises users by behaviour and not by biographical characteristics of their profile.

The work presented in this dissertation does not deal with the following conditions:

1. We assume that the malicious users are those that Twitter has banned and presented in the datasets available at the TTC. Therefore, we will not create new malicious user profiles or decide which users are malicious.

In Figure 1.1, the relationship between the hypothesis, the objectives and each of the research questions (RQ) can be seen. The research questions will be addressed in the following chapters.

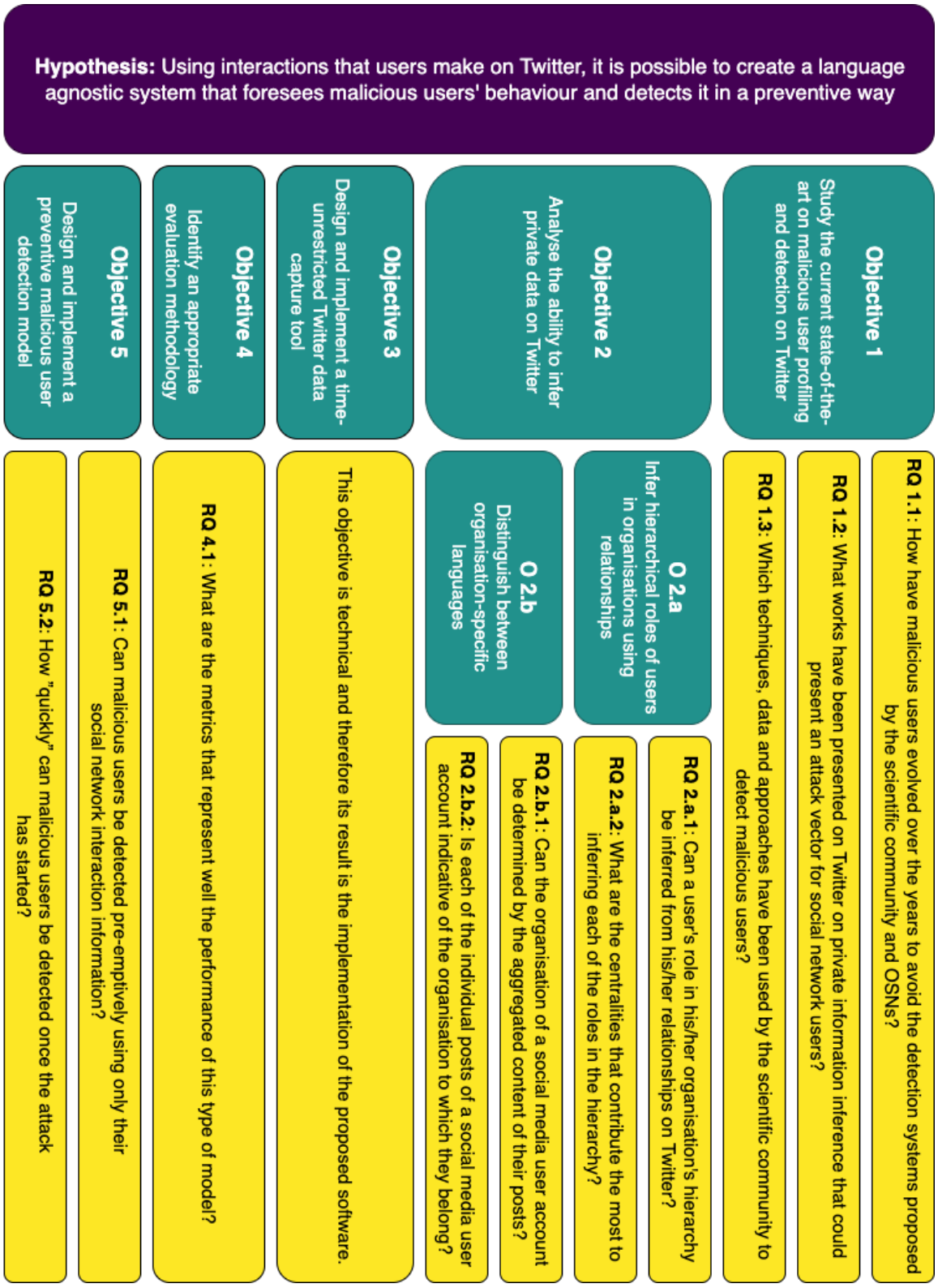

Figure 1.1: Summary of the hypothesis, objectives and research.

1.3 Research Methodology

The following strategy (see Figure 1.2) has been followed in order to accomplish the hypothesis and objectives presented in Section 1.2.

1. **Exploratory phase**: explore the literature related to the research field in order to build a solid theoretical framework on which support the rest of the work. Even though the task of revising the literature is presented as an isolated task, it is clear that it is an incremental and continuous process and consequently it will be done throughout the entire project.

2. **Definition of the validation scenario and conceptual test**: after the first phase of the study of the state-of-the-art, and knowing the limitations and advantages of the research proposal, a first version of the scenario in which the project will be developed will be defined. Even though it is expected to evolve during the investigation, having defined a strong initial framework will help steer the research towards the analysis that form the basis of it.

3. **Specification and design of the solution**: at this stage, the necessary requirements to solve the initial starting point will be specified, and the solution that can achieve the best results will be designed. To do so, it is essential to continue gathering information on the state-of-the-art, so that any innovation in the field will be considered in the design if it were relevant.

4. **Design of the test and evaluation**: after the design, the testing and evaluation system under which the solution will be assessed will be determined, verifying its validity and applicability in real environments.

5. **Development of a functional prototype, dissemination and writing the PhD dissertation**: the final task of the research will focus all efforts on the development of a demonstrator or functional prototype to justify each of the above tasks. In addition, the results obtained are expected to be innovative and of great significance for both academic

and social fields. Ultimately, this thesis will be finished and refined for its later submission and defence.

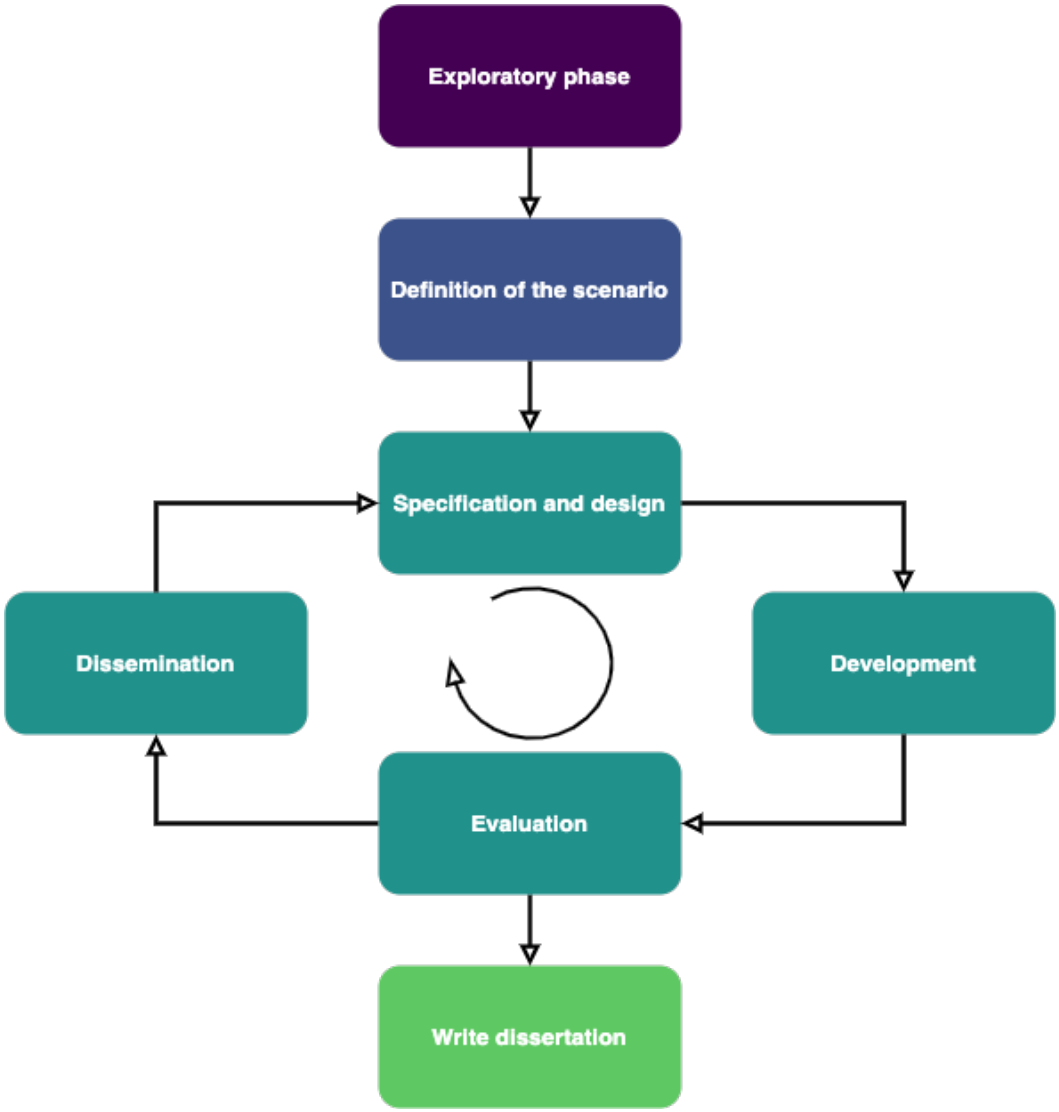

Figure 1.2: Schema of the research methodology employed in this dissertation

1.4 Main Contributions

This dissertation makes the following novel scientific contributions:

- An analysis of the vulnerabilities that organisations may have on Twitter based on the psychological principle of homophily. The analysis, as well as the methodology employed in it, are depicted in Chapter 3. This analysis is divided into two sub-contributions:
  - A generalisable method capable of deducing the roles of members of organisations using their Twitter relationships and the features extracted from the graph of their organisations. An extensive analysis of the node centralities is also presented as they represent the main features of the inference model. This method outperforms the previous state-of-the-art and is exported to be used in Twitter[1].
  - A novel study of tweets from members of five different organisations to prove the existence of specific languages on Twitter. The analysis introduces an ablation study of the different elements present in tweets (hashtags, mentions and link or Uniform Resource Locators (URLs)). It also introduces the differences that exist when classifying between isolated tweets and entire tweet timelines of users[2].
- A Deep Learning model that leverages Twitter interactions between users to foresee which users will become malicious after a specific set of actions[3]. The data employed for its training and testing is extracted from the TTC.
- A new methodology has been proposed to train and test malicious user detection models in a preventive way. The proposed methodology uses the data respecting the temporality in which they were originally created.

[1] https://github.com/rubensancor/posgcn
[2] https://github.com/rubensancor/pytorch_organisation_lang
[3] https://github.com/rubensancor/Mondrian

- A dataset containing 3 graphs composed of users from 3 organisations united by their relationships on Twitter. This dataset was employed to analyse the inference of roles in organisations using relationships between Twitter users (Sánchez-Corcuera, 2022c). The description of the dataset, as well as its usage, are described in Section 5.3.

- A dataset containing 41,545,828 tweets from members of five different organisations (Sánchez-Corcuera, 2022b). This dataset was employed to analyse if they use specific languages differentiable among organisations, as shown in Section 5.4.

- A dataset containing 596,221 tweets from legit users related to malicious users from three state-backed operations detected by Twitter and offered publicly in TTC. This dataset was used to train and test the model capable of preventively detecting malicious users (Sánchez-Corcuera, 2022a). The dataset and the motivation to create it are explained in Section 5.5.

The following technical contributions can be found in this dissertation:

- Twitter data-gathering wrapper: A software tool that enhances Twitter's API by adding features from other software to increase its data collection capabilities[1]. The most important features of this tool are explained in Section 5.2.

1.5 Thesis Outline

The thesis is structured into six chapters.

Chapter 1 is the current chapter. It presents the context and motivations of the research as well as the hypothesis, objectives and scope. A research methodology is proposed to achieve those objectives and validate the hypothesis. Finally, an outline of the dissertation is presented.

Chapter 2 describes the state of the art related to the dissertation. It introduces the methodology used to retrieve the articles and their metadata.

[1] https://github.com/rubensancor/TPL

It also presents the studies done through the decade of work against malicious users to profile them and the detection models developed by the research community.

Chapter 3 focuses on the open source intelligence available on Twitter and its use to infer private information from communities. It also presents the work related to hierarchical role inference and the existence of organisation-specific languages on Twitter.

Chapter 4 presents the malicious user foreseeing model. This chapter introduces the source of the malicious user data verified by Twitter and the system used to supplement it with legitimate users. In addition, the model used for this task and the details of its implementation are presented.

Chapter 5 presents the evaluation methodology followed for the models proposed in the previous chapters. Subsequently, the evaluation of the different models proposed in the dissertation is presented with a discussion.

Chapter 6 summarises this research work's main findings and contributions and shows the related future lines of work.

What you're seeing and what you're reading is not what's happening.

Donald Trump

CHAPTER

2

Related work

The arms race between social bots and researchers has existed since the beginning of the OSN. Derived from the realisation that these types of users represent a severe problem for people, researchers in the area started to profile and identify them by their features. Bots' creators responded to this by evolving their malicious users to make them less identifiable to user profiling systems. By losing the ability to be profiled, this new version of malicious users carried out attacks on Twitter that had worldwide repercussions, such as in the 2016 US elections, which gave rise to much research on the topic (Le et al., 2019; Bovet and Makse, 2019; Grinberg et al., 2019; Bessi and Ferrara, 2016; Enli, 2017). In addition, new forms of attack vectors on Twitter were discovered, such as private information inference by leveraging open source intelligence (Valverde-Rebaza et al., 2018; Zarrinkalam et al., 2018; Preoţiuc-Pietro and Ungar, 2018).

As a result, researchers in the field began to automate and modernise the models that detect malicious users by adding new features and methods to identify them. However, malicious users have reached a point where it is practically impossible to differentiate them by using profile features (Cresci,

2020), so new methods, such as graphs, or the exploration of posting patterns, have emerged.

Due to the large amount of information generated in the last decade in this area, we felt it necessary to present a review of the work done to give a background on how malicious users have evolved and what models have the scientific community employed to counteract them. To do so, we have followed the methodology presented in Section 2.1, which aims to answer the research questions proposed in the same section.

Therefore, Section 2.1 presents the methodology employed and the research questions proposed for this chapter. Then, Section 2.2 will explain the profiling processes conducted on malicious users through their different evolutions. Subsequently, we will introduce in Section 2.3 the previous works done in private information inference. Finally, the efforts done by the scientific community in automated malicious user detection are presented in Section 2.4. Finally, Section 2.5 will present the summary and conclusions of the chapter.

2.1 Methodology employed

For writing the related work chapter of this dissertation, we identified the need to review what progress has been made so far in the area of malicious user detection on Twitter and identify gaps where progress can be made. To this end, we propose three questions that, through their answers, provide the necessary context for understanding the proposed contribution and allow us to identify the gaps. These are the proposed questions:

- **RQ 1.1** *How have malicious users evolved over the years to avoid the detection systems proposed by the scientific community and OSNs?*
- **RQ 1.2** *What works have been presented on Twitter on private information inference that could present an attack vector for social network users?*
- **RQ 1.3** *Which techniques, data and approaches have been used by the scientific community to detect malicious users?*

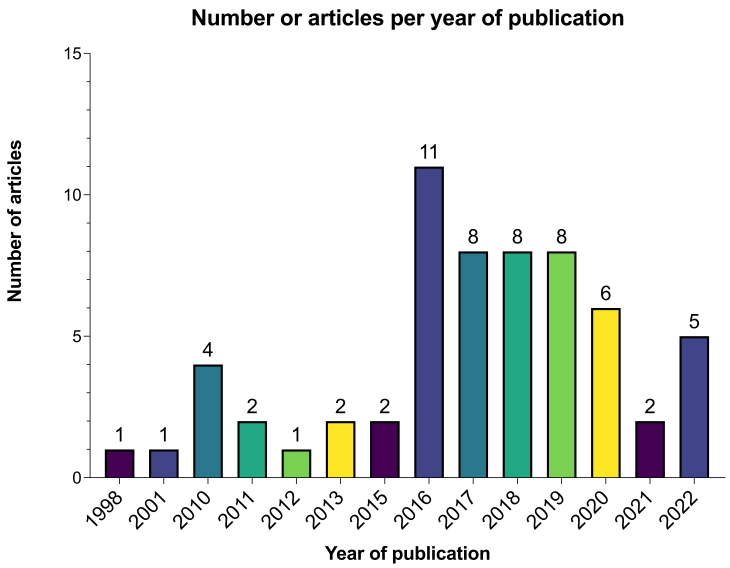

Figure 2.1: Number of articles per year of publication.

These questions give meaning to the following sections of the chapter, which aim to provide the reader with the necessary background to understand the topic of the dissertation. Furthermore, answering the questions allow us to envision the gap in which the advances presented in this dissertation fit. To answer the questions proposed above, we have conducted a state-of-the-art review of different aspects and tasks in the area.

To write the related work chapter of this dissertation, we reviewed 61 articles with publication dates from 1998 to 2022. As shown in Figure 2.1, there is a peak in the number of articles consulted from 2016, the year in which the scandal in the USA presidential election took place. From that year onwards, the number of articles related to detecting malicious users grew, making much progress in the area. Regarding the publishers, we consulted articles from 15 different publishers, as seen in Table 2.1. Finally, we consulted articles from organisations in all the continents except Africa following this distribution: North America (32), Europe (18), Asia (6), South America (4) and Oceania (1). Figure 2.2 shows this information in a more visual way.

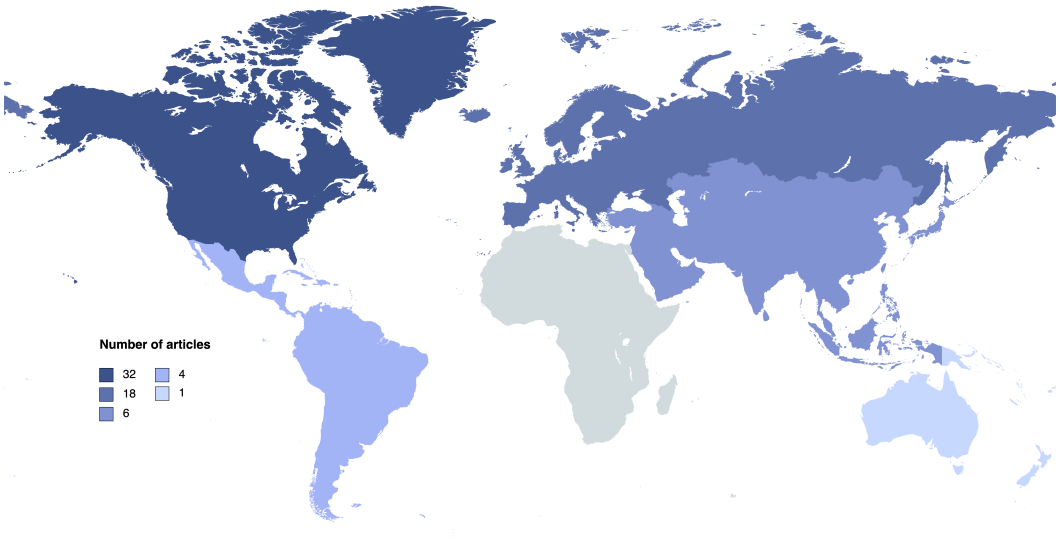

Figure 2.2: Number of articles per continent of the first author's affiliation organisation.

Publisher	Num. articles
ACM	14
IEEE	14
Springer	7
Elsevier	6
AAAI Press	5
arXiv	3
First Monday Editorial Group	2
ICLR	2
SAGE	2
ACL	1
American Public Health Association	1
ANNUAL REVIEWS	1
ASAC	1
MDPI	1
Usenix	1

Table 2.1: Number of articles per publisher

2.2 Profiling malicious users for a decade

Since the creation of Online Social Networks, the competition between malicious users and researchers trying to fight them has been constant. The first attempts to detect this type of user date back to January 2010 (Yardi et al., 2010). Since then, many researchers have tried to create systems capable of identifying automated users attempting to carry out malicious actions on Twitter. However, malicious accounts have constantly been evolving, making these models obsolete for the new versions.

To detect these malicious accounts or bots, researchers in the area make use of different features that help their model classify them from legit users. Therefore, we will present and expand the taxonomy presented in (Gilani et al., 2017) to divide the characteristics researchers in the field have used to characterise or profile malicious users. In the following list, the different categories and the approach offered in the state-of-the-art will be explained:

- **Content generation:** Content generation groups together all the content creation done on Twitter by these users. This includes tweets, retweets, mentions, replies and the content that can be included in a tweet, from URLs to the media in the form of photos or videos.

- **Content metadata:** Content metadata includes tweet information about the time when it is posted, the device used to post it or the location from where it is posted.

- **Content consumption:** Content consumption refers to the consumption of content bot accounts make. To measure this, we can only look for a bot account's number of liked tweets.

- **Content popularity:** Content consumption refers to the interactions that tweets produced by bots accounts create, i.e. how many likes or retweets the tweets produced. It also includes how many lists include a bot account, indicating its popularity or attention.

- **Network topology:** The topology of a user's network refers to friends and followers and their relationships. Also, the graphs and subgraphs these relationships create and their centralities.

- **User profile:** In this category, we group all the features related to users' profiles, such as the image, biography, name or location.

The first attempt to identify spammers on Twitter was published by Yardi et al. in (Yardi et al., 2010). The authors conducted an experiment to identify spammers on Twitter and subsequently concluded that some features differentiate them from legit users. Among those features were the date of creation of the account, the tweeting frequency and several regarding the topology of the networks of those users.

Furthermore, Chu et al. (Chu et al., 2012) also studied to characterise bots and cyborgs by their behaviours. They collected over half a million accounts with more than 40 million statuses. In their study, they analysed entropy measures in the timing of publishing statuses, the spam content of the tweets, information about the user profile and the device he/she uses to tweet. Using these features, the authors also proposed a classification system that obtained 96% of true positives.

As we can see, in these works, the authors used a little set of features to characterise spammers or bots. Therefore we can assume that the creation of these malicious users is not very complex and that their attacks are based on spam or simple techniques. Table 2.2 has a summary of malicious users' changes over time in their different versions and related articles.

Several years later, Yang et al. (Yang et al., 2013) published an article demonstrating that Twitter spammers and bots were evolving to evade the methods that identify them. In their study, the authors identify the evasion tactics followed by malicious users to evade the models that classify them. In addition, the authors selected the most common features that the models used to characterise malicious bots and enhanced them to follow those evasion tactics and identify the new version of spammers.

Following the work of (Yang et al., 2013), other authors (Zheng et al., 2015; Chen et al., 2015) improved their bot detection models to adapt to new

versions of them and improve their detection capabilities. In both articles, the authors developed models that obtained outstanding results in detecting the new wave of bots and presented tables in which they ordered the importance of each feature in detecting these users.

This new version of bots has evolved, and the models require more features to profile them. These new evasion methods created by bots include the obtention of more followers and the increase of tweeting ratio together with the mix of normal and spam statuses. However, with the evasion techniques, researchers in the area also identified new features that distinguished them from legitimate users.

In 2016, after the political election of the USA, numerous articles and papers were published analysing the work of some Twitter accounts dedicated to publishing and distributing disinformation in the run-up to the vote (Bessi and Ferrara, 2016; Bovet and Makse, 2019). However, although analyses were done and it was confirmed that bots or malicious accounts had the potential to affect countries' political processes, such users infected many more public debates (Bastos and Mercea, 2019; Mora-Cantallops et al., 2019; Pastor-Galindo et al., 2020; Rosenberg et al., 2020; Broniatowski et al., 2018).

After the attacks during the USA elections, some researchers noted that the paradigm on Twitter for malicious users was changing (Ferrara et al., 2016; Zhang et al., 2016b; Cresci et al., 2017). The authors of (Zhang et al., 2016b) were the first to demonstrate the existence of a social botnet under the control of a botmaster. Bots in this botnet conducted coordinated actions and mimicked legit users to reduce the possibility of being identified as individuals. This discovery denotes two evolutions from previous bots: the introduction of a botmaster coordinating the attacks and movements of the botnet and the addition of new features and behaviours to mimic legit users.

Ferrara et al. (Ferrara et al., 2016) identified that bots were being used to induct ideas in political discourses, manipulate the stock market or spread misinformation. The authors also indicate that, in some of the analysed attacks, these new malicious users were coordinated and obeyed the commands of a master that created the strategy followed in their attack. Finally, the au-

thors of the article state that due to the constant evolution of these malicious users, feature-based models have become obsolete to detect them.

However, the authors of (Cresci et al., 2017) confirmed the existence of a new wave of Twitter Spambots by checking that the state-of-the-art models could not detect them. This article confirms the statements previously done in (Ferrara et al., 2016; Zhang et al., 2016b). The authors confirmed that the new bots were coordinated and emphasised that new models for detecting them should look at user groups and their actions rather than individuals and their characteristics because of the level of mimicking with legit users these new bots reached.

Finally, and reaching the current times (2022), malicious users have once again been modernised to be undetectable by the models created by the scientific community. New technologies such as DeepFakes (Narayan et al., 2022) or language models (Sharma, 2022) have contributed to this evolution as they allow both the images used by users in their profiles and the language they use to be identical to that used by legitimate users. This is why feature-based models have become obsolete even though they report good results in datasets of suspended users. Moreover, attacks are nowadays not only perpetrated by users partially controlled by algorithms but also by real users who align with the objectives of the attack.

In addition to the features identified by researchers that profile these users and their evolution, the name the research community has used to refer to them has also evolved. Different names have been used depending on how they are managed (bots or cyborgs), the area in which they operate (political or antivaccine) or the type of attack they conduct (spam bots or astroturfs). However, during the evolution and even detecting that algorithms did not always control those users, the scientific community has not abandoned the bot word. Thus, we believe that the name that groups all these users regardless of their features is **malicious users** because of their malicious nature and objective.

In conclusion, as explained in (Cresci, 2020) by Cresci and as seen in both the table and the summary, the bots and malicious users have evolved over the years passing through different stages and becoming more sophisticated.

Furthermore, the emergence of technologies such as deepfakes or language models have made bots perfect themselves to the point of mimicking legitimate users and making them inseparable. Thus, the emergence of these complex bots makes the task of detecting them incomplete, and other tasks, such as detecting malicious users, regardless of how they are managed or identifying coordinated actions with malicious intent, arise from it.

2.3 **Private information inference**

Social networks offer the capacity for sharing information that can be used by other people with good or bad intentions. Adversarial information retrieval has been a security concern in social networks since their inception. Since the lastest trends given in OSNs, such as fake news on political matters (Grinberg et al., 2019; Bovet and Makse, 2019; Bastos and Mercea, 2019) and the rise of social bots (Ferrara et al., 2016; Varol et al., 2017; Davis et al., 2016), the literature in this area has grown considerably. Private information gathered on Twitter may be used by malicious users in their attacks as information to mimic themselves or for the search for weaknesses. For example, authors of (Das et al., 2016) discovered that malicious users change the style of their posts, therefore, the language, to achieve different objectives. This work shows that language is an attack vector that malicious users may use to attack users on ONSs. The recovery of adverse information may be carried out using different methods; however, in this dissertation, we will present several works that focus on the inference of private information using the public information available in Online Social Networks.

As stated by Gong and Liu (2018), the inference attacks conducted in OSNs can be classified into two categories: *relation-based* (Fire and Puzis, 2016; Valverde-Rebaza et al., 2018; Chen et al., 2016) and *behaviour-based* (Zhang et al., 2016a; Gong and Liu, 2016; Fani et al., 2019; Zarrinkalam et al., 2018). Relation-based attacks, which we will call behaviour-based, use relationships and attributes to infer the desired information. These attacks are based on the premise called *homophily* (McPherson et al., 2001), which

Version	Evolution	Behaviour	Identifiable by its features	Articles related
v1	Users that can be easily distinguished from legit users that spam continuously on Twitter.	Individual	Yes	(Chu et al., 2012; Yardi et al., 2010; Lee et al., 2010; Benevenuto et al., 2010; Wang, 2010; Ratkiewicz et al., 2011)
v2	Malicious users create evasion techniques to avoid detection algorithms. These evasion techniques consist of features that help them to mimic legit users.	Individual	Yes	(Yang et al., 2013; Zheng et al., 2015; Chen et al., 2015)
v3	Malicious users started to gain influence and disrupt different processes of societies. They started coordinating with botmasters that led the attacks, and feature-based models became less effective in detecting them, becoming obsolete at some points.	Coordinated	Some of them	(Ferrara et al., 2016; Zhang et al., 2016b; Cresci et al., 2017)
v4	Malicious users became homogeneous with legit users by using new technologies such as DeepFakes or Language Models. Thus, feature-based systems are completely obsolete for detecting this new version of users.	Coordinated	No	(Narayan et al., 2022; Cresci, 2020; Sharma, 2022)

Table 2.2: Summary of the history of the evolution of malicious Twitter users with their characteristics and references to articles in which they have been profiled.

states that users are more attracted to other users with whom they share characteristics such as religious beliefs or organisation. Instead, behaviour-based attacks use activity from users and some information from their relationships to infer information such as nationality or hobbies.

Regarding currently performed works on information inference on OSNs, we have identified two categories: activity inference and personal attributes inference. Work that aims at inferring activities uses the information publicly available on social networks to deduce which activities are being performed by the target users. In work proposed by Noulas et al. (2013) they propose an activity inference methodology using data from the OSN Foursquare[1] combined along with data from a telecommunication provider in Spain. The authors propose a classification task in which the telecommunication data has to be associated with the Foursquare semantic labels. Similar work was proposed by Chaniotakis et al. (2017) in which a combination of public data from Twitter and Foursquare was used to infer the activities conducted by Londoners. Authors use a data enrichment method by adding information about location and Foursquare activity tagging into tweets produced by the target users to classify them into an activity taxonomy. Although these works are centred on activity inference and are mostly behaviour-based, the classification and data gathering methodologies used in them are similar to the ones we proposed.

Works proposed for the inference of attributes in social networks are more common than those of activity inference. Related to attribute inference, many works aim to infer several or a specific attribute from a target user employing supervised learning machine learning algorithms with information extracted from the OSN (Zarrinkalam et al., 2018; Zhang et al., 2016a; Chen et al., 2016; Mei et al., 2017; Mulders et al., 2019; Gong and Liu, 2016). Many researchers have studied this topic because of many organisations' interest and recent attacks on Online Social Networks. For this reason, many researchers attempted to generalise the attribute inference and proposed models that can classify every possible attribute for a user, for example, AttriInfer. Jia et al. (2017) proposed a model based on modelling the social network as a pairwise

[1] https://foursquare.com/

Markov Random Field. This approach is based on behaviour and relationship information to decide if a target user will have a specific attribute.

Finally, in this dissertation, we proposed two inference works related to hierarchical role inference and linguistic style inference, which suppose a novel task in this research area. Thus, we will introduce specific works related to these topics in order to offer some background about the works done in them. Regarding hierarchical role inference on OSNs, (Chen et al., 2016) published a paper in which they used a *relation-based* approach to infer the roles of Google employees in Google+ OSN. Authors proposed a Naive-Bayes-based model that uses network metrics such as Degree Centrality or Cluster Coefficient and neighbour information to infer users' roles with promising results. Fire and Puzis (2016) proposed using scraped information from Facebook accounts of community members of organisations to infer the leadership positions in the communities of those organisations using supervised learning classifiers. Unlike these studies, our work incorporates directed graphs and a selected set of centralities that add more information to the machine learning algorithms allowing them to improve results. On the other hand, our work aims to generalise the inference of roles to other organisations not yet seen by the algorithm. Also, regarding the work of Fire & Puzis, our approach includes three categories of employees instead of detecting the organisation's leader.

Finally, linguistic style inference is a novel task in adversarial information retrieval; however, linguistic style-based text categorisation is related. In linguistic style-based text categorisation, researchers try to classify texts according to their writing style; similarly, we try to classify tweets and users in different organisations. As a first approach, authors of (Argamon-Engelson et al., 1998) propose the classification of articles from 4 different newspapers and magazines to demonstrate that the style of the text was independent of the article's topic. Later works proposed categorising text by author and genre using intrinsic features from the analysed texts (Potha and Stamatatos, 2018; Stamatatos, 2017). It has also been tried to combine corpus extracted from different platforms, social networks and movie review sites in this case, to show that languages from different platforms can help classify the authors of the

posts more effectively in some cases (Fourkioti et al., 2019). Finally, authorship identification has also been employed in OSNs to help forensics in cases requiring authorship identification for some post (Rocha et al., 2016). Derived from these tasks, the PAN workshop[1] focused on shared tasks on digital text forensics and stylometry as an annual challenge on authorship identification. Participants of this shared task are invited to develop models to classify texts from authors in 5 different languages using lexical, syntactical, structural and content-specific features (Bellot et al., 2018). These works demonstrate that the author's style remains even when the topic of the text changes. Whereas they are not focused on determining group or organisation-specific language styles, they provide encouraging evidence to support our hypothesis that language differs across organisations.

2.4 Automated detection of malicious users

Since the early identification of malicious users on Twitter, it has been detected that their actions could have repercussions on some processes in societies and even in people's lives. This is why researchers, in addition, to profiling these users, set about creating models or automated systems capable of classifying them among legitimate Twitter users (Yardi et al., 2010). This work has been going on for more than ten years (Cresci, 2020), and many systems have been implemented in order to detect these users using different methodologies.

Previous works have presented taxonomies dividing the models developed for automatic bot detection depending on the data used to identify bots. For example, the work presented in (Ferrara et al., 2016) classifies the work into: graph-based detection, crowd-sourcing detection and feature-based detection. However, in (Feng et al., 2022b), which is more updated, they replace crowd-sourcing detection with text-based detection, as collaborative methods were shown to be not very efficient and effective. In our case, we will adopt the second taxonomy to label the models we present in this section. However, it should be noted that currently, most of the models use more than one of the techniques discussed above; thus, combining different methods and

[1] https://pan.webis.de/

information may create a more accurate representation and classify users in a better way.

- **Graph-based detection:** These systems model social network information in the form of a graph by representing users as nodes and the interactions or relationships between them as the edges of the graph. In this way, the models can extract information about the behaviour of users and the structures they generate in their relationships. In addition, the graphs provide information on various characteristics of the nodes through the centralities.
- **Feature-based detection:** These systems use the information in the user profiles and their metadata to represent the users. These methods seek to identify users who share many characteristics and may be programmatically generated users.
- **Text-based detection:** These methods employ Natural Language Processing (NLP) models to process the text of tweets produced by users to generate representations of them. These methods can extract similarities between users' ideas or speeches and assist in classifying those promoting malicious speech.

In addition to classifying models by their methodology, researchers have also divided them according to the information they use to classify them. In our case, we will use the categories of information explained in Section 2.2.

Finally, the last defining feature of malicious user detection models is the use of data to train their models. There are two types of approaches based on proposed by (Milon-Flores and Cordeiro, 2022). In the first approach, called forensic, the data is used without respecting its temporality; thus, data in the dataset can be fully used without limitations. In contrast, the preventive approach respects the data's temporality by defining a point in the dataset that represents the present time; data after that point in the dataset remains unseen for the model until the test.

The forensic approach allows models to collect more information about users after the attack has been completed. In subsequent attacks, the models

can have more information to stop the attackers. However, the preventive approach represents the reality of the social network more faithfully as it sets temporal limits on what can be analysed before classifying users. In this way, models can be generated that are able to detect malicious users in pre-attack stages and prevent legitimate users from them.

2.4.1 **Proposed models**

The proposal for detecting malicious users on Twitter that we present in this dissertation (Chapter 4) is based on detecting patterns of behaviour preemptively through user interactions on Twitter, which could be classified as a graph-based detection method. Researchers in the area have presented several models using graphs or interactions between users to classify bots. However, none of them has presented a preventive approach. This subsection will summarise the most famous graph-based detection models and include some important models from other categories.

Automated malicious user detection systems, since their inception more than a decade ago with the detection of spammers on Twitter, have shared the same characteristics: (i) individualised user detection, (ii) reliance on supervised methods using data from previous attacks, (iii) focus on bot detection and (iv) employ a forensic approach. However, the methodology and data used in each model have been different; therefore, different results have been obtained for the same benchmarks.

The most similar work to the one we present in this dissertation was conducted by Cresci et al. in (Cresci et al., 2016). In it, the authors propose to create a digital DNA for each Twitter user by modelling their actions as letters of the alphabet, for example, "C" for comments and "L" for likes. Biological DNA sequences are successions of characters indicating the order of nucleotides within DNA molecules. Therefore, as biological DNA is used to search for similarities between humans, digital DNA may identify similar users on Twitter by looking for similar patterns in the DNA chains. The authors obtained more than 0.9 F-score with this methodology in two test sets (1 is the best score).

Regarding graph-based bot detection, all the proposed models employ the structure users made through their connections on the OSN. However, they also employ other data, such as user features or text, to enrich the classification and obtain better results.

In BotRGCN (Feng et al., 2021b), the authors propose the implementation of a heterogeneous graph using user relationships and apply relational graph convolutional networks to learn user representation for the classification of bot users. As features, the authors employ the text information from the tweets and all the available information from user profiles. Graph Convolutional Networks (GCNs) (Welling and Kipf, 2016) were also employed alone for bot detection by aggregating the same features as in BotRGCN to the relationship graph of Twitter and learning the representation of each user. After, these representations are classified with an MLP. Graph Attention Network (GAT) proposed in (Veličković et al., 2018) is an evolution of the Graph Neural Network (GNN) in which authors introduce attention to give different importance to neighbours. They also employ a Multi-Layer Perceptron for the classification task. BGSRD (Guo et al., 2021) is a specific use of both GCN and GAT models for bot detection on Twitter. The authors of this system use the RoBERTa model (Liu et al., 2019) to leverage information about tweets and user profiles as nodes of the graph exploited by GCN and GAT models. Heterogeneous Graph Transformers (Hu et al., 2020; Feng et al., 2022a) also uses GNNs but in a heterogeneous way by employing two different modules for mutual attention and message passing. In the mutual attention module, the model calculates the attention score considering the edge type and the source and target users. The message-passing model adds information about the source user and the edge dependency to the message before it is passed. All these three models employ the same features as BotRGCN.

Finally, related to graph-based methods, we wanted to mention one of the longest-running works presented in the bot detection task, Botometer (Sayyadiharikandeh et al., 2020; Yang et al., 2022). Botometer is a model based on more than 1,200 features extracted from available metadata information from interaction patterns, user profiles and content on Twitter. It has a public

Reference	Methods	Information used	Training approach
(Beskow and Carley, 2018)	Features	CG, NP, UP	Forensic
(Abreu et al., 2020)	Features	CG, CC, UP	Forensic
(Cresci et al., 2016)	Graph	CG	Forensic
(Echeverría et al., 2018)	Features, Text	CG, CM, CC, CP, UP	Forensic
(Lee et al., 2011)	Features, Text	CG, NP, UP	Forensic
(Yang et al., 2022)	Features, Graph, Text	CG, CM, CC, CP, NP, UP	Forensic
(Feng et al., 2021b)	Features, Graph, Text	CC, NP, UP	Forensic
(Feng et al., 2022a)	Features, Graph, Text	NP	Forensic
(Guo et al., 2021)	Features, Graph, Text	CC, CM, NP UP	Forensic
Ours	**Graph, Text**	**CG, NP**	**Preventive**

Table 2.3: Summary of the models presented in Section 2.4 with their characteristics. Here are the complete names of the information used in the models: CG: Content generation, CM: Content metadata, CC: Content consumption, CP: Content popularity, NP: Network topology, UP: User profile

website where active Twitter users can be evaluated and given a score on how similar are they to bots.

Although it has been shown that including the graph structure or complex models in malicious user detection systems considerably improves their performance (Feng et al., 2022b), there are also simple models performing well using features or text information solely. For example, the system proposed by (Lee et al., 2011) obtained good results in actual benchmarks by using 18 features fed to a Random Forest (RF) algorithm. (Echeverría et al., 2018; Beskow and Carley, 2018) also employ a RF algorithm with features extracted from users, contents and timing information. Finally, (Abreu et al., 2020) experimented using a significantly reduced set of five features and four different classifying algorithms tested in several benchmarks obtaining good results.

In Table 2.3, we present a summary of the models presented above, indicating the methods, information and the approach used to train and classify malicious users on Twitter.

2.5 Summary and Conclusions

This chapter presents a review of the state-of-the-art of research area addressed throughout this dissertation. Section 2.1 poses questions that are

answered throughout the three sections in which the state-of-the-art is reviewed. By reviewing previous work, the background necessary for the reader to understand the decisions made throughout the dissertation is presented. In addition, this review identifies the needs that exist in the area of malicious user detection on Twitter.

To answer the questions posed in Section 2.1, we reviewed 61 articles from 15 different publishers and authors on five continents. By answering the questions in the different sections of the chapter, it is possible to identify the gaps that exist in the detection of malicious users and where a contribution can be made. For example, in Table 2.3 in Section 2.4 it can be seen how there are no preventive models. This is why we can conclude that the chapter answers the questions by informing the reader about the state of the research area and enables the identification of areas to contribute.

Why is it that when one man builds a wall, the next man immediately needs to know what's on the other side?

Tyrion Lannister - A Game of Thrones

CHAPTER

3

Exploration of the Open-Source Intelligence on Twitter

Social networks gather information users post that can be employed to profile them. Marketing companies aggregate information from public profiles of different social networks to discover users' likes and dislikes and create more detailed profiles. However, most users do not publish everything on social networks as there is some private information that they keep to themselves. Nevertheless, this information can be discovered using machine learning models dedicated to private information inference. Malicious users employ this technique to discover sensitive information that may lead to attack vectors against their victims, be they individuals or communities of users (Heatherly et al., 2012).

In the context of social networks, attack vectors are the security breaches that malicious users leverage to carry out their attacks. They make users and communities vulnerable as they are not always aware of their existence, as many are based on psychological biases such as homophily. This is why it is

of utmost importance to demonstrate how easy it is to generate these attack vectors using techniques such as private information inference.

For the inference of private information, malicious users make use of open-source intelligence. This intelligence is defined in its Wikipedia entry as "The collection and analysis of data gathered from open sources to produce actionable intelligence". Therefore, in these attacks, open source intelligence is composed of all publicly accessible content on social networks or internet portals that host this type of content. In the specific case of the studies presented in this chapter, this information is limited to data extracted from Twitter and LinkedIn.

This chapter presents two analyses in which we unveil attack vectors by developing and evaluating private data inference models. The first analysis presents a model that infers the hierarchy of an organisation's members. The second one demonstrates the existence of organisational-specific languages on Twitter.

The remaining of this chapter is structured as follows: Section 3.1 describes the organisational role inference model proposed. Section 3.2 presents the model for demonstrating the existence of organisational-specific languages and how they can identify their members. Finally, Section 3.3 concludes this chapter with a summary of the main obtained findings. Most of the results described in this chapter have been published as journal articles (Sánchez-Corcuera et al., 2021a,b).

3.1 Inferring organisational roles through member relationships

This section introduces the first experimentation conducted to highlight the existence of attack vectors already existing and demonstrated on Twitter. Specifically, in this section, we address the objective O2.a of the dissertation by presenting a methodology to infer the hierarchical roles of Twitter users in their organisations by using their relationships on the OSN.

3.1.1 Objective and motivation

As explained in the previous section, inference attacks can be classified as relation-based or behaviour-based, depending on the information used to infer private data. In this section, we propose a relation-based experiment that uses the information on the relationships between Twitter users to infer the hierarchy of the organisation to which they belong.

We believe that the hierarchy of an organisation can enable an attack vector for malicious users by revealing who are the people with higher positions within the organisation. For this reason, we propose two research questions that we tackle in this work:

- **RQ 2.a.1** Can a user's role in his/her organisation's hierarchy be inferred from his/her relationships on Twitter?
- **RQ 2.a.2** What are the centralities that contribute the most to inferring each of the roles in the hierarchy?

To tackle these questions, we modelled the user relationships as a graph where nodes represent users and the relationships between them, the axes of it. This way, we can use the information of the nodes' centralities as features for classifying users in one of the three proposed hierarchical roles. Centralities, which will be explained later, have been used previously to classify nodes due to the information they represent and the network in general (Diallo et al., 2016; Zhan et al., 2017). Furthermore, we also conducted experiments to analyse the amount of information each of the centralities provides for classifying the different roles in the hierarchy.

In this section, we will limit ourselves to explaining the methodology and the models employed for the analysis. Section 5.3 presents the evaluation and the data employed.

3.1.2 Node centralities as features

In graph theory and network analysis, centralities stratify nodes taking into account different characteristics of their relations with other nodes in the

graph. As different centralities model different information, they have been used to analyse simple information from the nodes or combined to model more complex behaviours or information within the network. As they contain information about the nodes, how they relate and the network's topology, we select them to be used as features for machine learning classifiers in the proposed task.

For the experiments conducted we grouped the centralities into three categories: *Local-based*, *Global-based* and *Katz-based*. Local-based centralities (*Degree, Indegree and Outdegree*) are based on the edges going from and to the target node. Instead, global-based centralities (*Closeness and Betweenness*) include relationships from all the nodes in the graph and the edges that point to and from the target node in their calculations. Finally, Katz-based centralities (*Eigenvector, HITS and PageRank*) are different implementations of the original Katz centrality (Katz, 1953) that is based on the eigenvalues and the degree values.

We will present now each of the centralities that we used for our approach with their normalised formulas:

- *Degree Centrality (C_D)*: This centrality measures the number of edges connected to a node, that is, the number of neighbour nodes a vertex has. In this centrality, the most important nodes are the ones that are connected to a large number of nodes. Also, this can be interpreted as a risk as the most connected node has the highest probability of receiving any element going through the graph, for example, a fake new. This centrality has been used to separate fraudsters from legitimate users of an online auction by analysing how fraudsters collude more with each other to increase the price of items (Bangcharoensap et al., 2015). However, as it can be an important centrality to find if a node is an influencer in the graph, it is not a decisive property as stated by (Romero et al., 2011; Dey et al., 2019).

  Degree centrality is defined as follows, where k_i is the degree (number of edges connected to a node) of node i, and N is the total number of nodes in the graph (Opsahl et al., 2010):

$$C_D(i) = \frac{k_i}{N-1} \tag{3.1}$$

- *In-degree Centrality (C_{In})*: This centrality is derived from degree centrality and it is only available on directed graphs. It measures the number of edges directed to a node. On Twitter, influencers or famous users have higher in-degree centralities as they have a high number of followers.

  In-degree centrality is defined as follows, where k_i is the number of edges directed to node i, and N is the total number of nodes in the graph (Opsahl et al., 2010).

$$C_{In}(i) = \frac{k_i}{N-1} \tag{3.2}$$

- *Out-degree Centrality (C_{Out})*: This centrality is derived from degree centrality and it is only available on directed graphs. It measures the number of edges directed from the selected node to the others. Users with high out-degree in a graph may be identified as the users that deliver most of the information for the network.

  Out-degree centrality is defined as follows, where k_i is the number of edges directed from node i, and N is the total number of nodes in the graph (Opsahl et al., 2010).

$$C_{Out}(i) = \frac{k_i}{N-1} \tag{3.3}$$

- *Closeness Centrality (C_c)*: This centrality represents how close the selected node is to the other nodes of the network. Therefore, nodes with the highest closeness centrality will be the ones that can reach the others with the least steps or jumps possible. In social networks, this centrality is important to identify influencers as the ones with the highest closeness centrality are those that can broadcast the information more efficiently. This centrality identifies nodes that can spread, control or

acquire information from the networks; thus, it has been used to map networks of terrorist cells (Krebs, 2002).

Closeness centrality is defined as follows where i is the starting node, j the target node and $d(i,j)$ is the distance between them (Newman, 2010).

$$C_c(i) = \frac{N-1}{\sum_j d(i,j)} \tag{3.4}$$

- *Betweenness Centrality (C_b)*: This centrality measures the number of shortest paths that pass through a node. Therefore, nodes that are more likely to be in a path between the other two nodes receive higher values. Nodes with high betweenness centrality are often referred to as bridges since those nodes act as bridges for information between nodes. Nodes with high betweenness centrality, being in the middle of the information flow, have a position similar to that of a man-in-the-middle, so growing in this centrality is something malicious users crave. Due to the urge to search for short paths in telecommunications, this centrality has been used to increase the efficiency of this scenario (Borgatti, 2005).

  Betweenness centrality is defined as follows where i, s and t are three different nodes, σ_{st} are the number of shortest paths from node s and t and $\sigma_{st}(i)$ the number of shortest paths from node s and t that pass through i (Newman, 2010).

$$C_b(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma st} \quad C_b^*(i) = \frac{2C_b(i)}{(n-1)(n-2)} \tag{3.5}$$

- *Eigenvector Centrality (C_e)*: This centrality measures the transitive influence that a node has in a network. This centrality assumes that relationships with high-scoring nodes contribute more to this score than relations with low-scoring nodes. A high eigenvector score means the node is connected with many nodes with high eigenvector scores. Eigenvector centrality can be seen as an extension of the degree of centrality

that rewards the nodes for every connection they have with other individuals in the network.

This centrality is defined as follows where λ is the eigenvalue and A the adjacency matrix (Newman, 2016). This centrality is calculated recursively by giving each node a default value of 1 and calculating the new ones after each iteration until it converges. It can be calculated in all graphs, but as explained in (Newman, 2010) in directed graphs, as the adjacency matrix is asymmetric, two eigenvectors are calculated, left and right eigenvectors. In most cases, the right eigenvector is used as it represents edges pointing towards the selected node, but this can present a problem when nodes have no edges pointing to them. This is fixed in Katz centrality (Katz, 1953) by giving an initial score to all the nodes regardless of their position.

$$C_e(i) = v_i = \frac{1}{\lambda}\sum_j A_{ij} v_j \tag{3.6}$$

- *PageRank Centrality (C_p)*: This centrality is similar to Eigenvector one but with the difference that the score obtained by having an edge with an important vertex is diluted among all the nodes that are connected to that vertex. For example, the score of connecting with an influencer on Twitter is divided among all the followers he/she has. This centrality and the eigenvector one are used to present recommendations to other accounts to Twitter users by using shared interests and shared connections (Gupta et al., 2013).

  PageRank centrality is defined as follows where α and β are constants, A is the adjacency matrix, k_j^{out} is the out-degree centrality and x_j the value of the centrality for the node j (Brin and Page, 1998). A problem can arise if the node has an out-degree centrality of 0. To solve this problem, all the nodes with no outgoing edges have the default value of $k_j^{out} = 1$.

$$C_p(i) = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta_i \tag{3.7}$$

- *HITS Centrality (C_H)*: This centrality, called *hyperlink-induced topic search* (HITS), was designed by (Kleinberg, 1999) for directed graphs only. HITS centrality assumes two types of nodes in a graph: *authorities* are nodes that hold information about a specific topic, and *hubs* are nodes that know the routes to reach the authorities in a graph. However, a unique node can be an authority and a hub; they are not exclusive. These metrics have been used to characterise networks of propaganda on Twitter identifying users who have an active role on the propaganda and those who spread it (Guarino et al., 2020).

  For the calculation of HITS centrality, two measures are computed separately: authority (x_i) and hubs (y_i). The equations are as follows, where A_{ij} is the connection from the initial node to the rest of the nodes that are connected with it and A_{ji} vice versa. As in PageRank centrality α and β are constants.

$$x_i = \alpha \sum_j A_{ij} y_j \tag{3.8}$$

$$y_i = \beta \sum_j A_{ji} x_j \tag{3.9}$$

3.1.3 Employed models

As we will explain in Section 5.3, we employed several classification algorithms to compare our approach with the current state-of-the-art for hierarchical role inference on Facebook proposed by (Fire and Puzis, 2016). These were the machine learning algorithms selected for the experimentation: *Nearest Neighbours* (kNN) (Altman, 1992), *Gaussian Naive Bayes* (NB) (Hand and Yu, 2001), *Decision Tree Classifier* (DTC), *Random Forest* (RF) (Shalev-

Shwartz and Ben-David, 2014) and *Multi-layer Perceptron* (MLP) (Rumelhart et al., 1985).

As the data employed for our objective evaluation is in graph form, we also employed a Deep Learning model dedicated to graphs called GCN proposed in (Welling and Kipf, 2016). This model was developed by Kipf et al. because of the great attention that other data formats (text and images) were receiving from deep learning models, and he wanted to apply the same methods (convolution) to graphs as they are used in a wide variety of applications, such as social networks or protein-interaction networks.

In GCNs, the main idea is to apply the convolution operation (Bahri et al., 2013) to the graph structure. The convolution operation refers to multiplying some input neurons with a set of weights called filters or kernels. Convolution applied to graphs makes the model learn the features of the nodes by inspecting their neighbouring nodes. For this, the adjacency matrix is used as it specifies the relationships between the nodes. Introducing this matrix in the forward pass allows the model to learn the nodes' representations through their connections to each other. For a more specific explanation, refer to the original publication (Welling and Kipf, 2016).

3.2 **The existence of organisation-specific languages on Twitter**

Continuing with the methods of inference of private information on Twitter, this section presents the methodology used to implement a model capable of differentiating between organisation-specific languages used by its members on Twitter. The methodology proposed in this section focuses on solving objective 2.b of the dissertation.

One of the strategies that malicious users employ to conduct influential and credible attacks against organisations is infiltrating them by creating fake users with compelling features. This technique aims to create trust bonds and links with members of the attacked organisation by leveraging humans' tendency to build homophilic connections (McPherson et al., 2001). Homophily refers to

humans' tendency to relate and bond with those similar to them according to characteristics such as personality, behaviour or taste; homophily can apply to a wide range of dimensions, including gender, age, organisational role or class. Therefore, malicious users could employ data extracted from users' profiles to create homophilic links with them.

For creating fake avatars, malicious users employ profile pictures of real people or images generated by Generative Adversarial Networks (GANs) (Choi et al., 2018; Karras et al., 2020), capable of generating reliable faces. They also use legit information in their biography and post tweets related to the topics discussed in the group they want to attack (Zannettou et al., 2019; Freitas et al., 2016). Using information related to the groups they want to attack, malicious users establish trust bonds and links that make the attack easier (Rathore et al., 2017; Jun et al., 2017).

In this second evaluation related to exposing the ease of finding attack vectors on Twitter, we hypothesise that members of organisations use a specific language that malicious users may copy to impersonate or create trust bonds to attack them on Twitter. These impersonation attacks are common on social networks such as Twitter and Instagram (Goga et al., 2015; Zarei et al., 2019b). On Twitter, the most famous case of impersonation is that of Warren Buffett[1], in which an account was created imitating his language and using his image and biography. On Instagram, on the other hand, impersonations are done by using photos of the target person to create an account with the same characteristics as the original ones and comment on posts to gain visibility and spread polluted content (Zarei et al., 2019a).

3.2.1 Objective and motivation

Therefore, motivated by the latest attacks and works done around security on social networks, we question whether the common language among the organisation members can present an attack vector for the organisation as a whole. To this end, we proposed an analysis to determine if organisations

[1] https://www.darkreading.com/analytics/anatomy-of-a-social-media-attack/a/d-id/1326680

and their members use specific and distinguishable languages to detect and prevent attacks against them using these techniques.

We hypothesise that it is possible to know from which organisation a user is by merely using the language of their tweets. This hypothesis, in turn, enables tackling our research question, which states that organisations have specific languages.

To conduct the study, the analysis is divided into two different tasks: (1) classification of tweets in isolation to predict the organisation its author belongs to (tweet-level), and (2) classification of users by organisation by using their entire timeline of tweets (user-level). First, we aim to demonstrate that individual members consistently use a specific language that links it to their organisations with the tweet-level task. Second, with the user-level task, we want to demonstrate that the specific language can also be inferred when using all of a user's tweets in aggregation. Based on this, we set forth two key research questions that we tackle in this work:

- **RQ 2.b.1** Can the organisation of a social media user account be determined by the aggregated content of their posts?
- **RQ 2.b.2** Is each of the individual posts of a social media user account indicative of the organisation to which they belong?

To tackle these two research questions, we create a dataset composed of members of five different organisations. For each of these organisations, we identified user accounts for members with social media presence and retrieved Twitter profiles and timelines of tweets for each. We address the problem as a classification task, using two models that leverage linguistic features of tweets to classify them and compare their performances against a traditional machine learning algorithm. The prediction performance of the employed models needs to be above an established baseline to confirm RQ 2.b.1 and RQ 2.b.2. We complement our study by conducting an ablation study to analyse the elements a tweet may have (hashtags, mentions and URLs) and how they contribute to the classification.

3.2.2 Employed models

To answer the proposed research questions, we used several different models or algorithms in order to find the best results and reported the best model of the three. As a baseline for the evaluation, we used four algorithms not directly related to the language classification tasks; in this way, we were able to see how these algorithms classified the languages of the organisations and served as a barometer for the difficulty of the task as this was the first time it was performed. As a second algorithm, we have designed an ad-hoc model based on Convolution Neural Networks (CNN) (Collobert and Weston, 2008) specialised in the language for text processing with several layers of depth. Finally, we have used the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), the best-performing model in the state-of-the-art for text classification.

As mentioned above, we use four algorithms unrelated to NLP tasks to act as a baseline. Although we will later analyse RF as the best performing, the following models were used: kNN (Altman, 1992), MLP (Rumelhart et al., 1985), RF (Shalev-Shwartz and Ben-David, 2014), and Support Vector Classifier (SVC) (Cortes and Vapnik, 1995). To feed the text data into the classifier, we tokenised the tweets using the BERT-tokenizer, which will be explained later and ran the algorithm using the default parameters provided by SciKit (Pedregosa et al., 2011).

The second language-dependent model employed for this task is called Multi-CNN. We opted to use convolutions because they have proved to solve text classification tasks (Lee and Dernoncourt, 2016; Poria et al., 2016) and also those related explicitly to tweet classification (Severyn and Moschitti, 2015; Bilbao-Jayo and Almeida, 2018).

The model receives the tweets with a fixed size of D x 300 (blue layer in Figure 3.1), with being D the length of the most extended tweet in the batch and 300 being the size of the embeddings. Next, tweets are fed into the convolutional neural networks (orange layers in Figure 3.1). This network comprises two convolution layers with region sizes of 8 and 9, respectively. As stated by (Zhang and Wallace, 2017), using multiple filters is recommended to learn

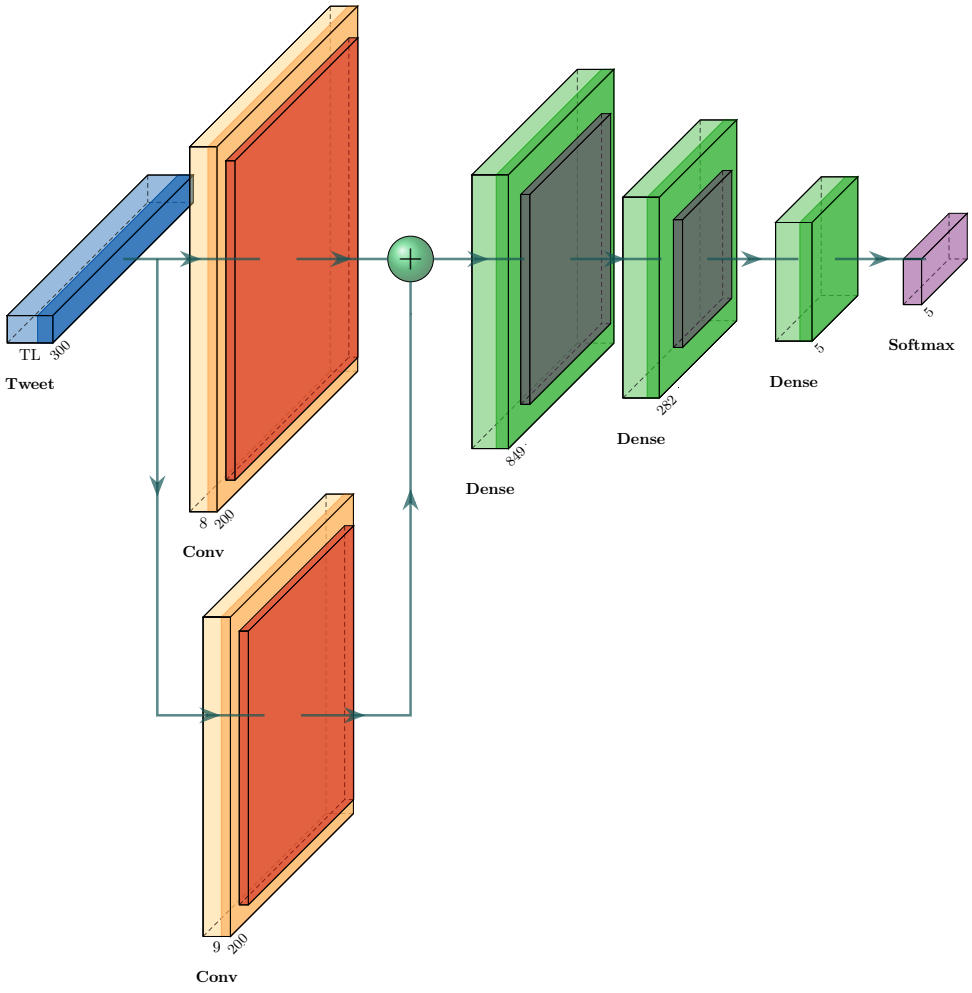

Figure 3.1: Plot of the Multi-CNN model.

complementary features about the analysed text. Therefore, each convolution produces 200 filters to which the activation function is applied, in this case, Rectified Linear Unit (ReLU). The result of each convolution is a vector of B x 200, being B, the number of tweets introduced to the convolution. Each convolution is followed by a ReLU function and a max-pooling layer to reduce the dimensionality of the data (Boureau et al., 2010) (red layers in Figure 3.1). We used max-pooling as it has been proved the best approach for natural language processing tasks (Zhang and Wallace, 2017). Finally, these two vectors produced by the convolutional layers are concatenated into a vector of B x 400 that contains the information extracted from both convolutions.

After the concatenation, the vector is fed two times to a pair composed of a fully connected layer with ReLU activation followed by a dropout (Srivastava et al., 2014) rate of 0.5374 to prevent the overfitting of the network (green and grey layers in Figure 3.1). The fully connected layers have 849 and 282 neurons, respectively. Finally, we applied a logarithmic softmax that computes the probability distribution of each label for the tweet (purple in Figure 3.1). We used the categorical cross-entropy loss function to train this model as it supports multi-class classifications and is the most common function for this type of problem. We used Adam, with the default parameters, for the optimiser; as it is the state-of-the-art for Deep Learning models as it slightly outperforms the others. The hyperparameter values and the size of the convolutions and the fully connected layers were decided after running a hyperparameter optimisation algorithm.

Finally, the last model employed for the evaluation was BERT. BERT has been one of the most important models for text classification since its creation by (Devlin et al., 2019) in 2018. When presented, it was the first model to achieve transfer learning successfully and achieved the best results for eleven NLP tasks in the state-of-the-art. It was conceived as a language model and pretrained using two document-level corpora, such as the BookCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words), in order to extract long contiguous sentences.

BERT's architecture was designed as a multi-layer bidirectional Transformer encoder and was implemented by stacking the encoder part of the

Transformer model proposed in (Vaswani et al., 2017). This model employs the self-attention mechanism to compute the representation of sequences and has been successfully used in several tasks such as question answering or language modelling.

The attention function proposed by the authors is called *Scaled Dot-Product Attention*; it receives three vectors called query, keys and values and computes the weight for each value using a compatibility function of the query with the corresponding key. Through this function, the model can identify the most important tokens for each token for the processed sequence. Furthermore, if this attention is stacked, it allows the model to attend information from different representation subspaces at different positions; this is what authors call Multi-Head Attention.

As we have stated, this model is bidirectional; thus, it can process the sequences from both left and right directions. Furthermore, it is not implemented using recurrence or convolution as the RNNs or LSTMs, so it has to employ what authors call the positional encodings. These encodings have relative or absolute position information about where are the tokens on the sequence so that the model can extract information about the order of the sequence.

BERT processes the input sentences using a tokenisation technique called WordPiece Model (WPM) (Wu et al., 2016) that converts words into tokens in a deterministic way using an existing vocabulary. In order to complete this, words are divided into subwords that can be converted back to the original ones by joining them together. In this way, we can decompose words that the model does not know into known words and extract some information from them. BERT has a vocabulary of 30,000 tokens that contains some special tokens for its own use, such as the special classification [CLS] which is always the first token and is used to hold the information of the sentence to classify.

In the original training of BERT, Devlin et al. employed two steps: *pre-training* and *fine-tuning*. In the pre-training step, the authors employ two unsupervised tasks instead of the traditional left-to-right or right-to-left language models used in previously presented models as in ELMo model (Peters et al., 2018). The first employed task is masked language model. In it, the

authors propose to mask 15% of all the input tokens of the dataset to predict them later using their context in both left and right directions. For the masking, the authors use the [MASK] token in which the model has to guess which of the words in all the existing vocabulary is masking. This allows to pre-train a bidirectional model but with a drawback, [MASK] token is only used on pre-training step and not in fine-tuning. To reduce the impact that this may have in the model, the authors propose that when the words are being masked they are replaced with [MASK] token 80% of the times, with a random token 10% of the times and leaving the word 10% of the time. The second task proposed for the pre-training step is next sentence prediction, where the understanding of the relationships between two sentences is studied. For this task, authors make sentence pairs `A` and `B` were 50% of the times `B` is the actual next sentence and 50% of the time is not. Thus, the model is trained in a binary way to classify if the `B` sentence is the one following or not the `A` sentence.

Devlin et al. proposed two different configurations for BERT: $BERT_{BASE}$ or $BERT_{LARGE}$. For our evaluation and due to hardware limitations, we employed the $BERT_{BASE}$ model as implemented in (Wolf et al., 2020). This configuration offers an encoder with 110M parameters that implements 12 transformer blocks, 12 self-attention heads and hidden layers of size 768. This model takes an input sequence of no more than 512 tokens. Finally, for this classification problem, BERT has a final layer to classify the possible labels; in this case, we employed a softmax function with the five possible labels.

BERT is already pretrained for text classification; however, fine-tuning must be conducted to adapt the model to a specific task. Therefore, we followed a strategy proposed in (Sun et al., 2019) consisting of unfreezing several hidden layers and training the model before running the tests. We conducted a hyperparameter optimisation experiment to find the best number of frozen layers to maintain in the model for this task. This experimentation leads us to unfreeze the last three layers and the classification layer, implemented through a softmax function, as this configuration obtained the best results for the proposed test. For the implementation of BERT, we used the PyTorch

version of the Transformer library developed by HuggingFaces (Wolf et al., 2020).

3.3 Summary and Conclusions

This chapter presents two analyses carried out to demonstrate that open source intelligence on Twitter can be leveraged to unveil attack vectors that may be used to harm communities. Section 3.1 presents a model capable of inferring the hierarchical roles of an organisation through the relationship between its employees on Twitter. Subsequently, Section 3.2 presents an analysis of the existence of a specific language on the part of organisations and the ease with which their members or tweets can be classified.

These analyses demonstrate the ease with which attack vectors can be generated to damage communities of users on Twitter. Due to this ability to compromise the security of Twitter users and together with all the attacks analysed in Chapter 2, it is intended to justify the subsequent chapter in which a preventive model for the detection of malicious users will be presented.

To ask why we fight is to ask why leaves fall - it is in their nature. Perhaps there is a better question: Why do we fight? To protect home and family, to preserve balance and bring harmony. For my kind the true question is: What is worth fighting for?

World of Warcraft

CHAPTER

4

Foreseeing Malicious Users Before they Attack

The constant manipulation attacks carried out against important events for society, such as elections for the presidency of a country, along with the attack vectors demonstrated by the scientific community are just one of the examples of the growing problem that bots and malicious users that plague social networks cause. The scientific community has proposed many models that classify legitimate users from bots or malicious users in response to these facts.

As explained in Chapter 2, most of these models' main problem is focused on detecting the so-called social bots and are mainly trained forensically. This training approach does not represent the reality of the social network because it uses all the data from each user to train the detection model. Thus, when applied using the Twitter feed as data, it will not work as accurately as in the previous tests because it will not have all the information about the attack as it had in the offline phase.

Furthermore, as stated in (Cresci, 2020) and also explained in Chapter 2, social bot accounts are becoming indistinguishable from legit ones due to the new advances in deepfakes or other deceiving algorithms. That is why

feature-based models are losing efficacy against new malicious user models, and researchers in the area have added new features to the detection models, such as the network topology of user relationships or activity patterns.

However, several researchers have pointed out that the field of malicious user detection has to change its focus toward detecting coordinated actions or groups/communities of malicious users. Since it is increasingly challenging to detect individuals, looking for patterns of coordination to detect malicious user communities can make the task of detection easier.

Twitter has helped this research through its Transparency Center (Twitter, 2020), where it provides information about attacks or different violations of Twitter's rules. In addition, this transparency centre has a specific section in which it publicly offers data on users who have taken part in manipulation attacks on the platform promoted by governments or states. This data used to train models can help to detect these types of actions or users.

In this chapter, we propose a pre-emptively trained model that is able to predict the actions that Twitter users will take in the future and thus characterises them as either malicious or legitimate. This way, the system can be used independently of the users' language during the attack. The model is trained and tested on data from state-backed operation attacks provided by the TTC with the intention that it is trained on users who are banned from Twitter and who perform coordinated attacks.

The chapter comprises various sections, starting from Section 4.1, which presents the objectives and research questions stated to prove the stated hypothesis. Then, Section 4.2 describes the existent datasets for bot detection created by the scientific community and the malicious user datasets employed for the proposed experiments, their origin and the reliability of the data. Section 4.3 describes the proposed model, its architecture. Finally, a summary and the chapter's conclusions are presented in Section 4.4. Some of the results described in this chapter have been published as a conference publication (Sánchez-Corcuera et al., 2022).

4.1 Objective and Motivation

Despite a substantial and increasing body of research in bot and malicious user detection (Cresci et al., 2015; Cresci, 2020), the vast majority of the work has been conducted with a forensic perspective (Yang et al., 2019, 2020; Cresci et al., 2016; Martinelli et al., 2019; Skorniakov et al., 2018; Mazza et al., 2019). The forensic perspective conducts experiments *a posteriori* using data from events that have concluded and have already been tackled. Previous approaches attempt to detect attacks post-hoc, whereas, in the work proposed in this chapter, we instead set out to introduce a preventive approach where we leverage user data preceding the events to predict them before they can finish their attacks. While the forensic approach allows the use of more comprehensive data, our proposed preventive scenario defines a realistic setting exploiting evolving data dynamics, and adapting to behavioural changes over time. We operationalise this preventive scenario by retrieving and expanding sets of malicious users published by Twitter under its TTC initiative, upon which we simulate the dynamic setting that exploits limited user histories preceding the events.

This chapter introduces a first-of-its-kind novel framework for the preventive detection of malicious users by leveraging information from the network topology, interaction with other users, and semantic features from URLs and hashtags in their tweets. We refer to malicious users who, regardless of how their accounts are managed, use them to carry out attacks and contaminate the network with fake information or idea induction attempts, for example. We build our model by exploiting the framework of JODIE, a dynamic graph network model proposed in (Kumar et al., 2019) capable of predicting the trajectory of user embeddings, thus, foreseeing if users will conduct malicious actions in the future. Our model exploits the benefits of JODIE by combining it with a classifier in charge of differentiating between foresaw embeddings for malicious or legit users. As far as we know, this is the first model trained and tested in a preventive way for this task.

To test the hypothesis and address the objective 5 of the dissertation, we have posed two key questions that we answer with this study:

- **RQ 5.1** Can malicious users be detected pre-emptively using only their social network interaction information?
- **RQ 5.2** How "quickly" can malicious users be detected once the attack has started?

We have set up several experimental processes to answer these questions, explained in Section 4.3, using a dataset created specifically for this task. This dataset is created by merging data publicly exposed by the Twitter TTC with data we have retrieved from legitimate Twitter users. In this way, we ensure that the social network itself verifies the users marked as malicious.

4.2 Data gathering and sources

In this section, we will introduce the different datasets the research community has collected over time regarding malicious users on Twitter and explain why we did not use them for our experiments. Subsequently, we will explain the source of our data and the origins and motivations of the Twitter Transparency Center. Finally, we will give details about the data we employed for our experiments and its statistics.

4.2.1 Malicious user datasets

During the several years that researchers have been detecting bots, content polluters or malicious users, the scientific community has collected a vast quantity of data with which they trained and validated their proposed detection models. Indiana University, through its Observatory on Social Media (OSoMe) (OSoMe, 2022), holds the most extensive repository of bot detection datasets. However, not all the datasets they offer are collected by them; thus, they are not offered in the same format or have the same objective or information. We will now introduce the most famous datasets created by the scientific community and compare them to justify why did we create our own dataset:

- *Caverlee-2011 (Lee et al., 2011):* For this dataset, authors create a honeypot from December 30, 2009 to August 2, 2010 were they collected 22,223 content polluters divided in four different types: (i) Duplicate spammers, (ii) Duplicate @ spammers, (iii) Malicious promoters and (iv) Friend infiltrators. (i) Duplicate spammers and (ii) Duplicate @ spammers post nearly identical tweets. However, the second ones mention random users in their tweets. (iii) Malicious promoters generate a good reputation to later use the account to promote online business, marketing or finance sites. Finally, (iv) Friend infiltrators seem to be legit users, but they abuse the reciprocity in following relationships. The authors also collected 19,276 legit users by gathering them randomly and monitoring them to verify their integrity. Finally, they gathered the number of follows and users they were following for each user over the time they were monitored.

- *Cresci-2015 (Cresci et al., 2015):* This dataset is composed by 1950 legit accounts and 1950 fake Twitter accounts. Legit accounts were gathered from the users that tweeted using the *#elezioni2013* hashtag except from politicians, parties, journalists and bloggers and a volunteer base of people that wanted to help in the academic research. However, the fake accounts were bought from three different fake Twitter account selling sites. The dataset comprises 2,631,730 tweets from legit users and 118,327 tweets from fake users.

- *Cresci-2017 (Cresci et al., 2016, 2017):* This dataset is composed by three types of Twitter accounts with the number of users gathered between parentheses: legit (3,474), social bots (991) and spam bots (464). Legit accounts were collected by asking them questions in natural language and manually verifying the received responses. Subsequently, social bots were gathered from being retweeters of the messages written by an Italian politician. However, these social bots showed genuine behaviour in their tweeting times and quantity and had their profiles created like real people, with images and biography. Finally, a set of spam bots promoting products from *Amazon* that interleaved the spam with

genuine tweets. Tweets from each type of user are divided as follows: legit (8,377,522), social bots (1,610,176) and spam bots (1,418,626).

- *Varol-2017 (Varol et al., 2017):* This dataset contains annotation of 2573 Twitter accounts producing English content collected by monitoring a Twitter stream for three months starting in October 2015. Collected users were filtered by restricting to those that had produced 90 tweets during the observation period and had 200 tweets in their accounts.

- *Gilani-2017 (Gilani et al., 2017):* Authors collected data from Twitter users for 30 days in April 2016, resulting in more than 65 million tweets and approximately 2.9 million unique accounts. They partitioned the data in four groups: G_{10M+}, G_{1M}, G_{100k}, G_{1k}. The number below the G means the number of followers that the users may have to pertain to each group. After collecting the users, authors conducted a human annotation task to identify users manually, ending with 1304 bot accounts and 1758 legit accounts taken from all four groups and nearly 400,000 tweets per user type.

- *Cresci-stock-2018 (Cresci et al., 2018, 2019a):* This dataset is based on a list of official cashtags of the most important US markets used as filters to obtain tweets that contain them. They collected more or less 9 million tweets and 2.5 million distinct users. In the dataset, authors found automated accounts that act coordinated to distort opinion about these markets or companies.

- *Midterm-2018 (Yang et al., 2020):* This dataset was collected during the U.S. midterm elections held in 2018, filtering tweets based on politics. Authors identified 8,092 legit accounts and spotted 42,446 bot accounts by correlations in their creation and tweeting timestamps.

- *Pronbots-2019 (Yang et al., 2019):* The dataset consists of a group of 21,963 bots that share scam sites and their tweets. Andy Patel shared it in his Github[1].

[1](github.com/r0zetta/pronbot2)

- *Celebrity-2019 (Yang et al., 2019):* This dataset contains 5,970 confirmed celebrity accounts with their tweets.

- *Vendor-purchased–2019 (Yang et al., 2019):* For these datasets, authors bought 1,088 fake followers and treated them as bot accounts, gathering also their tweets.

- *Botometer-feedback-2019 (Yang et al., 2019):* A set of 143 bots and 386 legit accounts gathered from the feedback accounts used in Botometer (Davis et al., 2016) and manually labelled.

- *Political-bots-2019 (Yang et al., 2019):* This dataset is a collection of 62 automated political accounts controlled by @rzazula that are now suspended. The dataset also contains the tweets of those users.

- *Cresci-rtbust-2019 (Mazza et al., 2019):* This is an Italian dataset of retweets shared in a two week time span. The dataset contains 1,591,865 original tweets related to nearly 10 million retweets.

- *Botwiki-2019 (Yang et al., 2020):* This dataset is based on the botwiki.org archive of self-identified bot account. This dataset contains 698 bot accounts with their tweets.

- *Verified-2019 (Yang et al., 2020):* This dataset contains users and tweets from verified accounts on Twitter. These accounts are certified by Twitter to be real people. It contains 11,987 verified accounts with their tweets.

- *Kaiser (Rauchfleisch and Kaiser, 2020):* Authors of this dataset detected manually 27 German bots and joined with 532 official accounts of German members of parliament together with 516 accounts of members of the 115th U.S Congress.

- *Astroturf (Sayyadiharikandeh et al., 2020):* A dataset of 505 hyperactive political bots pariticpating in follow trains and/or systematically deleting content.

- *Twibot-20 (Feng et al., 2021a):* This dataset comprises a comprehensive sample of the Twittersphere and aims to represent the current state of bots and legit accounts on this OSN. Authors gather the users by using a breadth-first search expanded from 40 different users from these four disciplines: politics, business, entertainment and sports. This set contains 229,573 users with more than 33 million tweets. Furthermore, this dataset is the only one that reports the following relationships between the users. However, although this dataset is the largest and most informative to date, it only reports 200 tweets per user at most, losing the temporality and dynamics of Twitter.

After analysing the datasets created by the scientific community for the detection of bots, we identified that most of them did not meet the requirements we needed to fulfil our objectives. In Table 4.1, we compare the users contained in each dataset and the requirements we set to meet the objectives previously presented. Furthermore, we explain these requirements in the following list:

1. Dataset must have both malicious and legit users in its data.
2. Malicious and legit users in the dataset should be related or gathered in the same period of time.
3. Dataset must contain the tweets of the users captured and respecting the temporality of the publications (respecting Twitter's broadcasting rules).
4. User labels, especially malicious ones, must be corroborated by Twitter by suspending accounts or marking users with the automated user label.
5. Dataset should contain malicious users independently, and how they are managed, this includes bot, cyborg and human accounts.

Due to these reasons, we decided to create a new dataset that contains official data of suspended Twitter users involved in state-backed coordinated operations and detected by the social network itself. Furthermore, we collected

Dataset name	Malicious users	Legit users	Users related	Timeline format	Annotated by Twitter	Malicious users
Ours	**1594**	**2736**	✓	✓	✓	✓
Caverlee-2011 (Lee et al., 2011)	15,483	14,833	✓	✓		
Cresci-2015 (Cresci et al., 2015)	1,950	1,950		✓		
Cresci-2017 (Cresci et al., 2017)	7,049	2,764		✓		
Varol-2017 (Varol et al., 2017)	733	1,495	✓			
Gilani-2017 (Gilani et al., 2017)	1,090	1,413		✓		
Cresci-stock-2018 (Cresci et al., 2018)	7,102	6,174	✓	✓		
Midterm-2018 (Yang et al., 2019)	42,446	8,092	✓			
Pronbots-2019 (Yang et al., 2019)	17,882	-	✓	✓		
Celebrity-2019 (Yang et al., 2019)	-	5,918	✓	✓	✓	
Vendor-purchased-2019 (Yang et al., 2019)	1,087	-		✓		
Botometer-feedback-2019 (Yang et al., 2019)	143	380		✓		
Political-bots-2019 (Yang et al., 2019)	62	-	✓	✓		
Cresci-rt-bust-2019 (Mazza et al., 2019)	353	340	✓	✓		
Botwiki-2019 (Yang et al., 2020)	698	-		✓		
Verified-2019 (Yang et al., 2020)	-	1,987		✓	✓	
Kaiser (Rauchfleisch and Kaiser, 2020)	27	1048	✓		✓~	
Astroturf (Sayyadiharikandeh et al., 2020)	505	-				
Twibot-20 (Feng et al., 2021a)	6,589	5,237	✓~			

Table 4.1: Comparison of bot detection datasets created by the scientific community and ours with the proposed requirements for our task.

data from legit users related to the malicious ones to create a complete dataset of legit and malicious users in the same circumstances. For it, we employed the data of the Twitter Transparency Center, which we will explain in the next section.

4.2.2 **Twitter Transparency Center**

Twitter created the transparency report in 2012 to inform users about the pressures governments put on social network regarding censorship or requests for information/removal of certain content. Subsequently, due to the need for a site to upload the information from the transparency reports, the announcements on new measures and their progress on transparency, they turned their transparency report into a website called the Twitter Transparency Center (TTC). This website contains several sections in which Twitter publishes information on transparency in three categories: (i) Legal Requests, (ii) Twitter Rules and (iii) Security & Integrity. In the first category, information is posted about legal requests made to Twitter for removing content at the government's request or for non-compliance with the Digital Millennium Copyright Act (DMCA), for example. The Twitter Rules section groups actions taken against accounts that violate the rules and Terms of Service of Twitter, such as the manipulation of information or spam or false information about COVID-19. Finally, the Security & Integrity section publishes the actions that Twitter takes to increase the protection and authenticity of the social network, such as the protection of accounts or information operations data, which we have focused on.

Information Operations are state-linked information operations that conduct inauthentic influence campaigns carried out by several users, sometimes with the knowledge and consent of the country's government. Since 2018, Twitter has been publishing the data of the operations detected so that the academia or journalists can use it to learn and investigate them and, for example, create systems that can prevent these types of attacks. Every few months, Twitter publishes the operations that have occurred since the last

publication in a new blog published in the TTC. These posts specify the operations detected, their targets and the number of users suspended as a result. In addition, along with each blog post, Twitter publishes a dataset with the details of the accounts involved in the operations and all the tweets involved in them.

In the state-backed operation datasets, Twitter anonymises the information of lesser-known users (with less than 5,000 followers) to minimise the impact or damage that can be caused to the compromised accounts in case of false positives in the attack. However, researchers specialising in Twitter can request a fully unhashed version of the data to conduct experiments as long as a data licence is signed, the data storage location is secured, and the data is used for research purposes only. In our case, we have followed the measures imposed by Twitter to be able to use this unhashed version of the data.

As of June 2022, Twitter's transparency portal published 42 of such coordinated actions. Due to the large number of actions, their diverse typologies and the costly process of collecting data from legitimate users, we decided to select three different actions. The countries of origin of the accounts are the following: China, Iran and Russia. The motives behind the actions and the actions sizes are different in each country:

- *China (July 2019):* Hong Kong held in 2019 a massive protests movement called Anti-Extradition Law Amendment Bill Movement or 2019 Honk Kong Protests were a series of demonstrations were held in response to the introduction of an extradition law. Twitter detected 936 accounts from within the People's Republic of China (PRC) with the intent of causing political unrest in Hong Kong, including discrediting the legitimacy and political stances of the protest movement on the ground. As Twitter is blocked in PRC, these accounts employed VPN or specific unblocked IP addresses from China. Twitter claims that these accounts are a small part of a network of more than 200,000 accounts dedicated to this type of action.

- *Iran (June 2019):* Twitter discovered 1,666 accounts associated with the Iranian government. Those accounts tweeted nearly 2 million times

about global news with a bias that benefited Iran's diplomatic and geo-strategic view. These accounts were suspended as platform manipulation is a violation of Twitter Rules.

- *Russia (May 2020):* 1,152 accounts were detected associated with Current Policy, a media website engaging in state-backed political propaganda within Russia. These accounts were cross-posting and amplifying content in an inauthentic, coordinated manner for political ends. These activities included promoting the United Russia party and attacking political dissidents.

Through manual inspection of the datasets, we observed that even though prior research has generally referred as bots to the accounts that carry out the attacks and the contamination of the network, the posts from the TTC reveal that not all the banned users in information operations are bots or fake users. Thus, we propose to group these users as malicious users and develop classifying systems that tag them regardless of how their accounts are managed. This is the main reason we group such users as malicious users throughout this dissertation.

We also noticed that in some cases, the number of accounts indicated in the Twitter transparency centre blog posts did not match the number of accounts provided in the datasets. In other cases, the number of detected/suspended accounts was the same, but not all of them had tweeted in the reported data. Thus, in this dissertation, we report the number of users and tweets that were available in the downloaded unhashed datasets.

	China	**Iran**	**Russia**
#Legit nodes	192,434 (344)	96,762 (684)	123,488 (1,527)
#Malicious nodes	191	389	974
#Edges	2,029,418	1,868,433	1,142,663

Table 4.2: Node and edge statistics of the datasets. The number in brackets refers to the number of nodes as being creators of interactions and therefore classified.

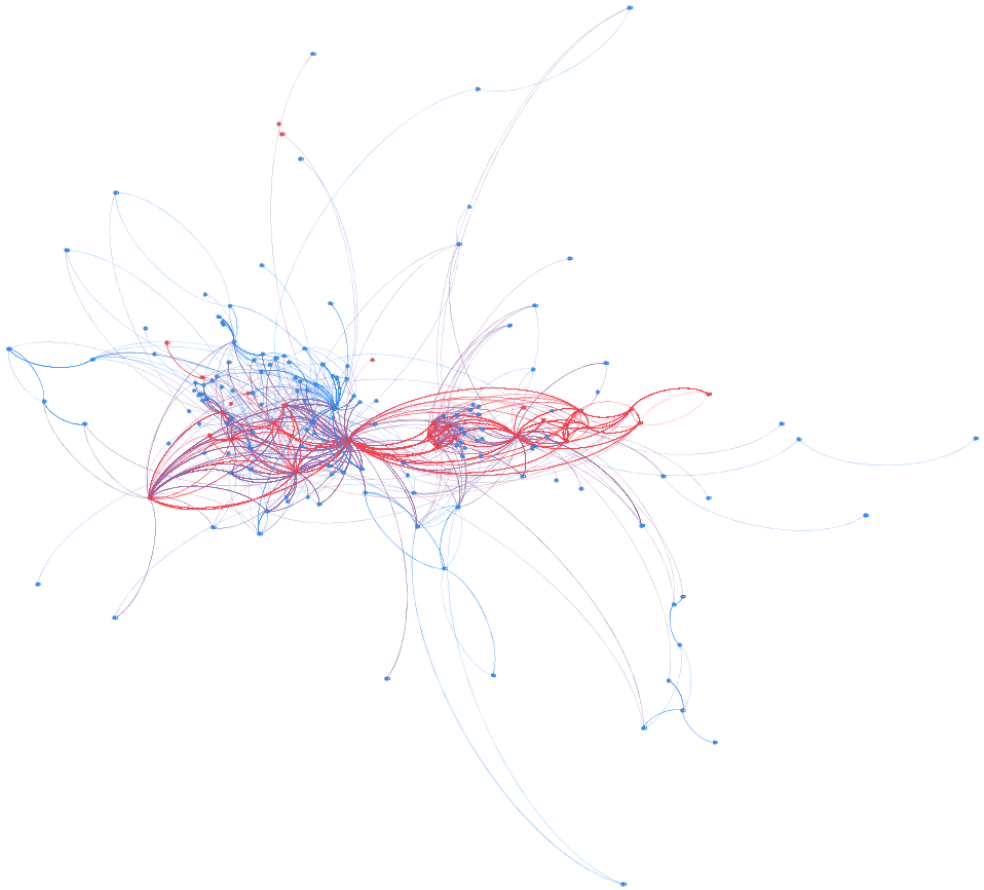

Figure 4.1: Sample from the Iran dataset graph. In red malicious user nodes and blue legit users nodes. The edges have the colour of the node they originate from. Purple edges symbolise one or more axes from each connected node.

For the collection of legit users, we decided to select users who had had a connection with the malicious users but had not been flagged by Twitter as such. Many of the attacks carried out on Twitter use hashtag-based strategies such as *hashtag flooding* or *hashtag hijacking* (Jain et al., 2015) to carry out their main attack. Therefore, we decided to use malicious users' most commonly used hashtags to harvest legitimate users. For selecting those hashtags in each attack, we grouped all the hashtags used in the tweets from the offered

User type	Interaction type	China	Iran	Russia
Legit	Mention	26.32%	19.97%	3.70%
	Reply	19.45%	25.46%	7.35%
	RT	2.00%	2.28%	0.79%
	Plain tweets	52.23%	52.28%	88.16%
	Total tweets	321,812	557,750	1,489,737
Malicious	Mention	6.46%	3.12%	1.34%
	Reply	16.40%	2.51%	13.96%
	RT	39.12%	38.33%	12.08%
	Plain tweets	38.02%	56.04%	72.62%
	Total tweets	1,707,606	1,310,683	3,441,965

Table 4.3: Breakdown of tweets in the datasets by type of interaction and user.

dataset and selected those that represent 80% of the hashtags from the dataset. In this way, the legitimate users would connect with the malicious users through those selected hashtags. To do so, we introduced the selected hashtags to the crawler tool presented in Section 5.2 and collected the information from the users who had tweeted with the same hashtags. Then, we used another crawler to get all the tweets posted by legitimate users 48 hours before and after the coordinated action. Due to the time needed for the crawler to collect tweets, particularly for prolific users with many tweets, we decided to limit the number per legitimate user to 1,000. The tweets from legit users are available here (Sánchez-Corcuera, 2022a).

Finally, we merged the data of malicious and legitimate users by creating a dynamic graph that used the interactions between users as edges and the users as nodes, as shown in Figure 4.1. We used the primary forms of user interaction on Twitter for the edges: RT, mention and reply. Since many of the attacks carried out on Twitter make use of misleading hashtags or URLs, we decided to use their encodings as features of the edges so we could leverage them with the foreseeing model. The statistics for each dataset are shown in Tables 4.2 and 4.3.

4.3 Proposed model

To carry out preventive detection of malicious users on Twitter, we decided to employ a model specialised in predicting the next interaction between users and items. For example, this model has been used to predict the next interaction between users and subreddits in Reddit's online social network. In addition to the interaction foreseeing model, we added a final classifier to classify which users will be potentially malicious in the future based on the foresaw interactions.

In order to perform embedding foreseeing and embedding classification, we have proposed a two-stage model as depicted in Figure 4.4. Each model is in charge of predicting the next interaction of users and classifying the users between malicious or legit users depending on their representation given by the previous model respectively. These models are explained below:

4.3.1 Embedding foreseeing model

This model is responsible for projecting users' embeddings to the desired time. To do this, we build on a model called Jodie (Kumar et al., 2019), initially designed for link prediction and node classification tasks. As this model is crucial for the prediction of the movements of malicious users we will explain its implementation and training details in the next section.

4.3.1.1 JODIE

The JODIE model was proposed by Kumar et al. to model sequential interaction between users and items in different domains such as e-commerce or social networks. In it, the authors employ representation learning by embedding both users and items in Euclidean space and analysing their trajectory through their interactions over time. For its implementation, the authors used two RNNs, each in charge of modifying the embeddings of users or items. Finally, the most crucial feature of the proposed model is the future modelling of the trajectory of users and items. For it, the authors created a projection operation that learns to estimate the embeddings at any time in the future.

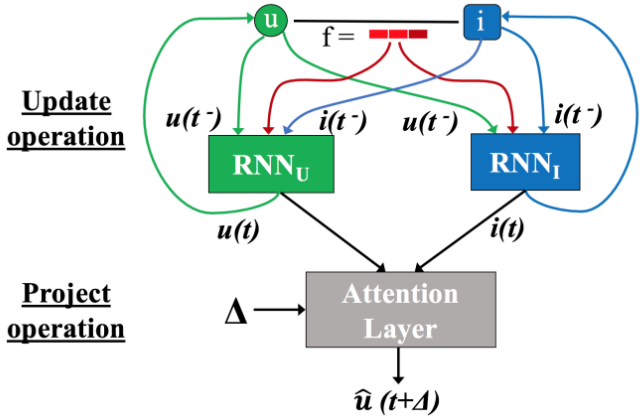

Figure 4.2: Update and project operations of JODIE model. Figure extracted from (Kumar et al., 2019) for explanatory purposes.

As this is the main model employed for the experiments conducted in this thesis chapter, we will explain its implementation. The symbols and nomenclature are extracted from the original article (Kumar et al., 2019) to ease the article's further reading if necessary.

JODIE is proposed as a method to learn embedding trajectories of users U and items I using an ordered sequence of temporal interactions between them. Each interaction happens at a specific time, and it may have an associated feature vector (e.g. the rating message of a user and an item in an online shop). Two embeddings are assigned for each user and item: static and dynamic embeddings. Static embeddings do not change over time and are in charge of expressing the long-term interest of users. Thus, the authors implemented these embeddings using one-hot vectors as they have proven to give similar results to more complex solutions. Subsequently, dynamic embeddings represent users and items at a specific time and change over time to model their behaviour and properties. The sequence of these dynamic embeddings is what authors call trajectory.

JODIE is proposed in two operations in charge of updating and projecting the embeddings after each interaction, as shown in Figure 4.2. We will now explain those operations more deeply:

- *Update operation*: In the update operation, a interaction $S = (u, i, t, f)$

between a user u and item i at a given time t, with its features f is used to generate the dynamic embeddings of that user and item at that specific time. The model uses two Recurrent Neural Networks (RNNs), called RNN_U and RNN_I; each one is in charge of updating the embeddings users or items, respectively. Authors designed these RNNs as mutually-recursive, i.e. when an interaction is processed at time t, the RNN_U updates the dynamic embedding of the user by using the previous embedding of that user $u(t^-)$ and item $i(t^-)$ right before the time t and regardless which has been its last interaction and the same for updating the items. The update operation for users and items is defined as:

$$\begin{aligned} \mathbf{u(t)} &= \sigma(W_1^u u(t^-) + W_2^u i(t^-) + W_3^u f + W_4^u \Delta_u \\ \mathbf{i(t)} &= \sigma(W_1^i i(t^-) + W_2^i u(t^-) + W_3^i f + W_4^i \Delta_i \end{aligned} \tag{4.1}$$

Where Δ denotes the time since the previous interaction of a user or item depending on its sub-index ($\Delta_u or \Delta_i$). f is the feature vector of the interaction and $W_1^u ... W_4^u$ and $W_1^i ... W_4^i$ are trainable matrices of RNN_u and RNN_i respectively. The authors justify using RNN instead of LSTMs, GRUs or other similar architectures because they experiment with them and obtained similar or worse results with those models that include more parameters.

- *Projection operation*: This operation is in charge of projecting the embedding of a user in the euclidean space to the desired time. With this operation, the model can predict the trajectory of a user and, thus, predict the next item with which that user will interact. The dynamic embedding of a user u at a specific time t plus the elapsed time is necessary for this operation. First, they convert the time to a time-context vector by passing it through a linear layer. Then, the projected embedding is obtained by conducting an element-wise product of that time-context vector with the previous embedding of the user. More formally:

$$\hat{u}(t+\Delta) = (1+w) * u(t) \tag{4.2}$$

Where $1+w$ represents a temporal attention vector to scale the embedding of the user in the past. Thus, when $\Delta = 0$ the time vector $w = 0$ and the projected embedding is the same as in the time t. The larger the value of Δ, the more difference between the original embedding and the one projected as the time difference is larger.

The authors of JODIE train the model to predict the next item a user will interact with. To train both update and projection operations, the authors decided to train the model using the projected embedding of the user $u(t+\Delta)$. One important decision made by Kumar et al. was to directly output a projected item embedding, $\tilde{j}(t+\Delta)$, instead of a probability of the interaction between the user and the items. Thus, when conducting the forward-pass of the prediction layer, the JODIE model outputs the predicted item embedding and the item with the closes embedding to the projected user embedding is returned as the next interaction.

Thus, JODIE model is trained to minimise the L_2 difference between the predicted item embedding $\tilde{j}(t+\Delta)$ and the real item embedding $[\bar{j}, j(t+\Delta^-]$ as the concatenation of the original embedding of item j and it embedding immediately before the time Δ. This difference is calculated as follows: $||\tilde{j}(t+\Delta) - [\bar{j}, j(t+\Delta^-]||_2$. For the prediction of the future item embedding authors propose to employ the projected user embedding $\hat{u}(t+\Delta)$ the embedding of the previous item that the user has interacted with $i(t+\Delta^-)$ at the prior time to the interaction Δ (that is what the superscript $-$ means). Finally, the static embeddings of both user ($\bar{u}$) and item ($\bar{i}$) are used in the fully connected layer used for the prediction as follows:

$$\tilde{j}(t+\Delta) = W_1\hat{u}(t+\Delta) + W_2\bar{u} + +W_3 i(t+\Delta^-) + W_4\bar{i} + B \tag{4.3}$$

where $W_1^u ... W_4^u$ and the bias B make the linear layer.

As mentioned before, JODIE is trained to minimise the L_2 distance between the predicted item and the ground truth item in the interactions. However,

the loss applied to the training phase includes two terms to prevent the dynamic embeddings of users and items vary too much and also these terms are scaled using λ_U and λ_I to ensure that losses are in the same range. The loss is calculated as follows:

$$\begin{aligned} Loss = & \sum_{(u,j,t,f)\in S} ||\tilde{j}(t+\Delta) - [\bar{j}, j(t+\Delta^-]||_2 \\ & + \lambda_U ||u(t) - u(t^-)||_2 \\ & + \lambda_I ||j(t) - j(t^-)||_2 \end{aligned} \tag{4.4}$$

The original model is trained using what authors call *t-batches*, special batches developed to maintain temporal dependencies between interactions. To maintain these temporal dependencies, the authors create these t-batches following these two requirements: (1) all interactions in a batch should be processed in parallel, and (2) the batches should be processed in increasing order to maintain the temporal ordering of the interactions. Thus, these two requirements may be summarised in that two interactions of the same batch do not share any common user or item. The authors prove that this batching system makes JODIE 9.2 times faster than its closest state-of-the-art model called *DeepCoevolve* (Dai et al., 2016).

Finally, Kumar et al. proposed to prove the performance of the model against six different models of the state-of-the-art of three different categories: (i) Deep recurrent recommender models, (ii) Dynamic co-evolution models and (iii) Temporal network embedding models. For the comparison, the authors propose five experiments to prove that they are outperformed in several tasks and characteristics such as runtime or robustness.

Unlike previous work by (Kumar et al., 2019), who used source and target users to predict future interactions, our model exploits Twitter interactions, enabling us to predict users' future actions. This selection of entities allows us to predict the user's strategy and classify him/her as a malicious user before he/she finishes his/her attack.

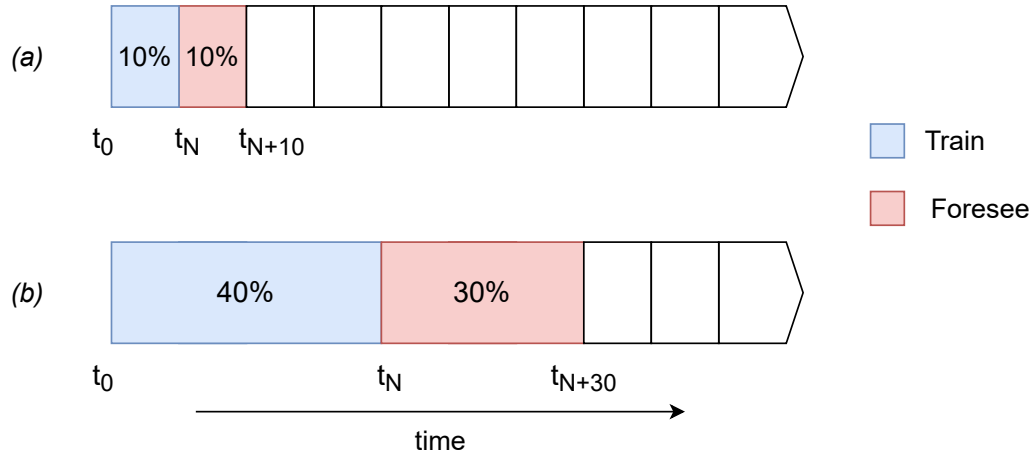

Figure 4.3: Examples for dataset splits used in our experiments. *(a:)* The model is trained with 10% of the data (t_N) and with a foreseeing size of 10% of the data (t_{N+10}) *(b:)* The model is trained with 40% of the data (t_N) and with a foreseeing size of 30% of the data (t_{N+30})

4.3.2 Embedding classification model

This model is in charge of classifying the embeddings of the users that the *Embedding foreseeing model* produces. This model is implemented using Random Forest (Breiman, 2001) classifier with 300 estimators from the Scikit-learn library (Pedregosa et al., 2011). Several algorithms were tested during the model selection study, and the one with the best results across the three datasets was selected. The results of the model selection study can be seen in Table 5.13.

4.3.3 Methodology

In this subsection, we will explain the methodology followed to train and test our model with the data extracted from the TTC. Figures 4.4 and 4.3 the whole methodology followed in our experiments and give details about how is the data divided for training and testing. We will now enumerate and explain the steps followed in it:

1. The dynamic multigraph created from the TTC dataset is processed and converted into a time-ordered series of interactions $(t_1, t_2 ... t_X)$ between

users. Each interaction has an origin, destination and a list of features. The features for these experiments consist of the tokenised versions of URLs and hashtags used in the tweets. In the leftmost part of Figure 4.4 the interactions are represented as edges between users and labelled as a_{RT} or a_m depending on the conducted action.

2. The training part of the dataset is defined. For this, we will select a moment in the dataset that will represent the present, and all the previous data will represent the past data; thus, the following interactions will be unseen to the model, which will be in charge of foreseeing them. For example, in Figure 4.3.a, the training part consists of the first 10% of the interactions. We will denote the final timestamp of the training part t_N. Interactions until t_N are fed to the *Embedding foreseeing model* to train the embedding projection layer to create user representations after foreseeing each user's future interactions.

3. The last timestamp used to foresee user interaction is decided, i.e. to what extent do we want to foresee the interactions in the future. In 4.3.a, the testing part will be the interactions after t_N until 20% of the total interactions in the dataset have taken place. That timestamp in the dataset will be called t_{N+10}. The model will now foresee the embeddings of each user that has an interaction in between those times.

4. Finally, the *Embedding foreseeing model* will produce the embedding of the users at t_{N+10}, and those will be passed to the *Embedding classification model*. This model makes a hundred random splits to train and classify the embeddings provided by the previous model.

As we can see, the model can be applied directly to Twitter data. The model would be trained with old data of the desired users up to the current time, then the users' movements would be predicted, and their future intentions could be classified.

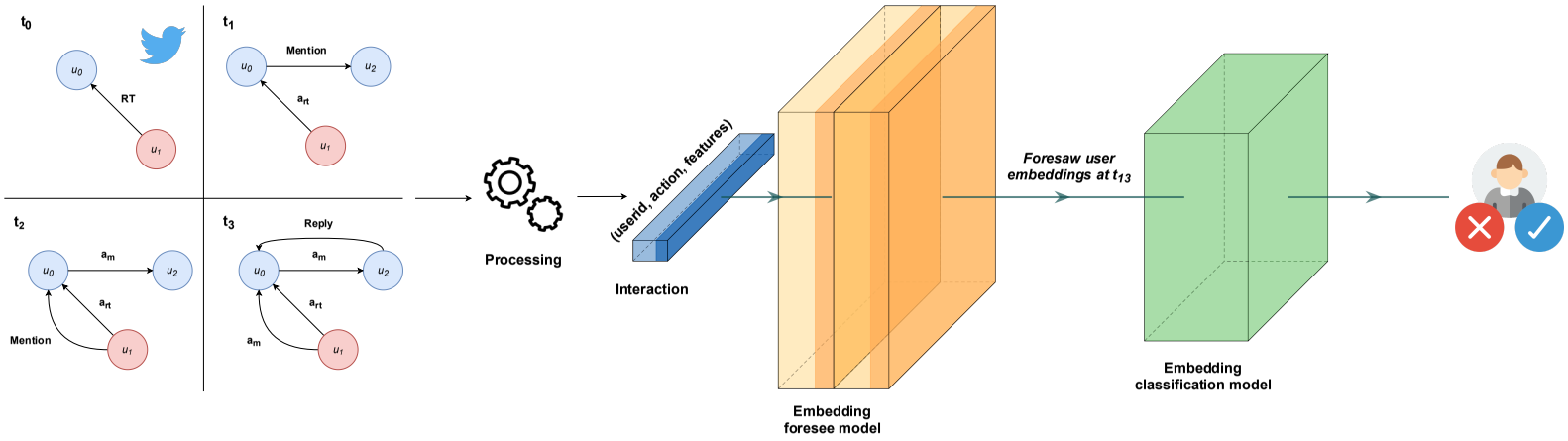

Figure 4.4: The methodology of our approach, from creating the dynamic multigraph to classifying the embeddings at the desired point of the dataset. In the leftmost part of the figure, we can see a representation of how the dynamic multigraph is created. In t_1 the u_1 makes a RT to the u_2 and the first edge between them is created. The actions follow one after the other, increasing the sub-index of t and thus creating the temporality of the graph. This dynamic graph will later be processed into a list of interactions sorted by date.

4.4 Summary and Conclusions

This chapter presents the most relevant work carried out in the thesis. A system of preventive detection of malicious users that works by leveraging the information of users' interactions to detect when they are carrying out attacks that Twitter considers harmful to the social network.

To do so, we proposed using data on attacks previously carried out on Twitter, which the social network offers researchers through its Transparency Portal. Subsequently, we collected information on legitimate users in order to create datasets with which to train a model capable of classifying them preventively. In other words, we have proposed a model that predicts the future interactions of users and then classifies whether the predicted actions will make the user malicious or not.

While we believe that such models are important because of the increasing number of attacks that are being carried out on social networks and that significantly impact everyday processes, they also raise certain ethical issues that need to be addressed. Such as freedom of expression or deciding what is an unbiased social network.

Furthermore, as a conclusion related to the model, we can highlight that the experiments proposed for the model are adequate to evaluate its performance in different scenarios and the results obtained, as shown in Chapter 1, are very good considering the complexity of the task. The model is able to detect malicious users in advance, so we believe in its viability. We indeed see a clear future work to make it generalised and to be able to deal with all types of attacks detected on Twitter.

Who knows? Have patience. Go where you must go, and hope!

Gandalf - Lord of The Rings

CHAPTER 5

Evaluation

The following chapter presents the evaluation methodology employed for the validation of the models proposed in the previous chapters and its obtained results. First, the metrics employed for validating the evaluation and their explanation are presented in Section 5.1. Section 5.2 presents the data collection methodology and the main features of the library implemented for it. In 5.3 shows the achieved results in the role inference task and a discussion about the obtained results. Subsequently, Section 5.4 offers the evaluation of the demonstration of the existence of organisational-specific languages. Then, the results and a discussion about the results obtained in the malicious user detection model are presented in Section 5.5. Finally, the conclusions are presented in a summary form in Section 5.6.

5.1 Evaluation Methodology

We employed accuracy and macro F-score as measures for our classifiers for the proposed evaluations, and the results are reported in the following sections. We employed these measures because they are the most popular ones when

evaluating a classifier's performance. Furthermore, these metrics are used by researchers in the tasks in which we conducted evaluations in this dissertation.

When evaluating the performance of a classifier, the points can be divided into four types depending on how the system has labelled them. Imagine we have an algorithm to classify points into relevant or non-relevant ones. The relevant points classified as relevant are called True Positives (TP). However, the points the system has incorrectly classified as relevant are called False Positives (FP). The rest of the points not classified as relevant are also divided into two categories. The relevant points not detected by the system are called False Negatives (FN), and the correctly non-relevant points not labelled as relevant are called True Negatives (TN). The most common metrics, such as accuracy (shown in the equation below), contain these metrics in their calculations. Accuracy is one of the most widespread measures to check the viability of a classifier in the face of a problem.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

However, this can be a problem when our datasets are imbalanced and the classifier always uses the majority label. For example, in a dataset containing 100 examples from two different classes, where 80 of them are from the same class, if our classifier always chooses the majority class, it will have an accuracy of 80% even if it does not classify correctly. Therefore, it is necessary to accompany this metric with others that consider these limitations, for example, F-score, which we will explain below.

F-score or F-measure (F_1), known as the harmonic mean of the precision and recall, is a measure of a test's accuracy. Precision is the number of true positive examples divided by the number of labels classified as positive by the classifier, including those that are not correct. Recall, instead, is the number of true positive examples divided by the number of all the labels that should be identified as positive. This measure aims to combine the values of precision and recall in one unique value as a valid test measure for classifiers. The best score achievable with F-score is 1, and the worst is 0. The equations below present how these metrics are presented more formally.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The previously presented examples and metrics are focused on binary classification, and, as explained in previous chapters, some of our evaluation processes have multi-label classifications. Thus, reporting recall or precision measure would not represent the system's performance as it would show the model's performance predicting one specific class.

However, the scientific community has developed alternatives for averaging evaluation metrics in order to be used for multi-label classification systems (Opitz and Burst, 2019). There are two different averaged F-scores: micro and macro. In this dissertation, we will focus on macro F-score as the most spread averaging method and is the one that is used in the state-of-the-art for some of the presented works; thus, we employed it in our evaluations.

Macro F-score gives the same importance to each label/class. Unlike F-score, the metric will be good if the dataset is unbalanced and the system performs well in the most common class. However, using macro F-score in the same situation, the performance achieved will be low for the model that only performs well in the most common class but poorly in the classes with few examples.

5.2 **Data collection methodology**

In Chapters 3 and 4, analyses have been proposed to validate the objectives and hypothesis of this dissertation. In order to evaluate such experimental processes, large amounts of data are needed. This section will introduce the methodology used for data collection and its tools. We will start by discussing the Twitter API, its limitations, and the alternatives proposed to overcome them.

5.2.1 Twitter API

Twitter API was released two months after the launch date of the online social network and became a reference for public REST APIs. Twitter offered a set of methods that exposed the intern workings of the platform so developers could use them for their applications or develop new Twitter user interfaces. The API also offered full access to tweets and content published by users on the OSN. As no limitations were imposed and the social network generated large amounts of data, the API attracted interest among developers, people in academia and journalists. Due to this interest and attraction of users, during the first year, the API obtained a strong growth and reputation.

However, in 2012, Twitter introduced the most damaging change for developers, the rate limitation per endpoint (Sippey, 2012). These limitations restrict the access per application to the methods exposed by the API, causing many applications to be abandoned due to their unusability. The alternative to the limits imposed was to employ the expensive Enterprise version of the API that Twitter offered with higher rates and fewer limitations. Nevertheless, some applications did not need that much access to the platform and could not expend the money that accesses cost.

In July 2020, Twitter released the second version of the API, including new features that were missing in the previous versions and introducing new access levels (essential, elevated, elevated+ and academic/research). These new levels are more dedicated to the use and collection of data; however, the premium and enterprise licenses also extend the rates of use of most of the methods available in the API. The essential is the basic access level that is given at the creation of the developer account. The elevated level is given by sending an application form to Twitter in which the reason the extended access is needed has to be specified and the use of the data explained. Finally, the academic/research access level is given by demonstrating belonging to an academic environment in which data from social networks is analysed and the publication of several works with data from Twitter. More information about the requirements and access levels is available on the APIs web (Twitter, 2022).

Twitter's API has many limitations concerning data or the usability of the platform's functionalities. However, we will focus on the limitations which have caused problems in completing the analyses proposed in this thesis. The most common one is the rate limit of 900 requests every 15 minutes, which slows information retrieval. However, the most limiting constraint is that only the most recent 3,200 tweets can be gathered from each user, together with the restriction of 500,000 tweets a user can pull from Twitter. All the rate limits, together with the tweet caps, can be consulted in their developer portal (Twitter, 2022).

Due to the rate limits, tweet caps and the time restriction of tweets on Twitter, we decided to explore new methods to gather data from Twitter, avoiding, when necessary, using the API. Thus, in the next section, we will explain which was our approach to data gathering.

5.2.2 **Tweets Processing Library (TPL)**

Since the implementation of restrictions on the use of the Twitter API, developers and researchers using data from the social network have implemented many methods to retrieve data bypassing those barriers. Although many libraries are available on Github and already implemented with many capabilities, we focused especially on Twint (OSINT, 2019) for all the relations it has with our requirements.

Twint, or Twitter Intelligence Tool, was created as a scrapping tool that gathers information from Twitter users without using the official API. The benefits it offers are the ability to bypass the limit of tweets retrievable by users imposed by Twitter and the possibility of using it anonymously, that is, without using a Twitter account. These features are essential for us as they allow us to gather the data we need with quick library deployment. Besides scrapping Twitter information from users, Twint also offers the visualisation of relationships between users as a graph. This feature is also relevant for us as we modelled our datasets as graphs in some of the experiments proposed in this dissertation. However, Twint developers discontinued its development

in 2019. Thus, most of its features are not functioning, as Twitter has been changing its webpage to prevent scrappers from working.

Therefore, motivated by the work done previously by other developers, specially Twint, and the restrictions imposed by Twitter in their API, we decided to develop a Twitter Processing Library (TPL)[1]. In TPL, we unified all the methods employed to gather or process the data from Twitter to create the datasets used in our experiments. For some of the methods, we had to bypass the restrictions of the official API, but we also employed the API when it was faster than other methods. Furthermore, having all the methods applicable to our datasets in a single library allows us to work in a unified way and create them with the same format. In the following list, we introduce and explain some of the most important methods implemented in the TPL:

- **Information gathering:** We used Tweepy to extract some information using the official Twitter API.
- **Scrapping:** We implemented a scrapper with the initial intention of scrapping tweets from searches on Twitter. However, we adapted it to scrap information from specific users.
- **Tweets processing:** We implemented several methods to process the tweets to extract the necessary information from them or their authors. Also, we implemented methods to delete some information from the tweets, such as emojis or unusual characters. Finally, we also implemented the necessary methods to process the tweets into the final datasets.
- **Graph generation:** We implemented a method to implement graphs from relationships between Twitter users. In the created graphs, nodes are represented by users and edges by relationships or interactions between them.

Besides the functionalities listed before, we also implemented an algorithm dedicated to retrieving Twitter users related to organisations or specific movements based on the one proposed by (Fire and Puzis, 2016). This algorithm

[1] https://github.com/rubensancor/TPL

Algorithm 1 Organisational Mining Algorithm based on the one proposed in (Fire and Puzis, 2016). The default priority is set to 30.

Input: A set of seed Twitter Usernames (S) of the organisation's employees and a set of words related to the target organisation, N

Output: A set of Twitter profiles with twitterid, name, username, biography, followers and followings

Organisational Miner() :

$Q \leftarrow Priority-Queue()$
$\forall_{Username \epsilon S}, Q.Enqueue(Username:30)$
$Crawled \leftarrow \emptyset$
$NonRelatedUsers \leftarrow 0$
while $(Q \neq \emptyset$ & $max(Q.Priority) \neq 1$ & $NonRelatedUsers < 1000)$ **do**
  $Username \leftarrow Q.Dequeue()$
  $Page \leftarrow DownloadTwitterProfileData(Username)$
  $Crawled \leftarrow Crawled.append(Username)$
  if $NinPage$ **then**
    $Connections \leftarrow ExtractConnFromTwitter()$
    $Connections \leftarrow Connections - Crawled$
    for $(Connection \epsilon (Connections \cap Q))$ **do**
      $Increasepriority(Connection)$
    end for
    for $(Connection \epsilon (Connections - Q))$ **do**
      $Q.Enqueue(Connection, Priority:1)$
    end for
    $CollectedPages.append(Page)$
  else
    $NonRelatedUsers++$
  end if
end while
return $CollectedPages$

was implemented to gather users of the first two studies proposed in this dissertation. The algorithm receives a list of seed users related to one specific organisation and a set of keywords related to that specific organisation or its name. It checks the followers and followings of the selected users and whether they have one of the keywords in their user profiles. Subsequently, the algorithm continues with this process with the new users it has retrieved. The limitation of the proposed algorithm is that users must have one of the words set as keywords in their user profiles. The pseudocode of the specific algorithm is presented in the Algorithm 1

5.3 Evaluation of Role Inference Model

In this section, we propose that hierarchical roles of employees in organisations may be inferred using their relationships on Twitter. To validate the first part of the second objective of the dissertation, we posed two research questions (RQ 2.a.1 and RQ 2.a.2). This section will introduce the experimental setup proposed to answer both questions, the dataset used, and a discussion about the obtained results.

5.3.1 Experimental setup

To answer the research questions posed for the completion of the second objective of the dissertation, we have conducted an evaluation directed to analyse how the topology of the OSN helps to infer members' roles in the hierarchy of their organisation. In this section, we will introduce the experimentation conducted and its setup.

The first research question (RQ 2.a.1) aims to clarify if the hierarchical roles of Twitter users in their organisations can be inferred by using their relationships on the social network. To answer this question, we propose to evaluate a model's performance in classifying users in their roles. As explained in Section 3.1, this question was answered for Facebook OSN by (Fire and Puzis, 2016). Therefore, we conducted the same experimentation process for both approaches to compare how different relation-based inference models

perform. To make both approaches comparable, we employed the same classifying algorithms as in the current state-of-the-art. The classifiers used for the evaluation are the following ones: *Nearest Neighbours* (kNN) (Altman, 1992), *Gaussian Naive Bayes* (NB) (Hand and Yu, 2001), *Decision Tree Classifier* (DTC), *Random Forest* (RF) (Shalev-Shwartz and Ben-David, 2014). Furthermore, we also added *Multi-layer Perceptron* (MLP) (Rumelhart et al., 1985) and a Deep Learning model dedicated to graphs called Graph Convolutional Network (GCN).

The second research question (RQ 2.a.2) aims to clarify which centralities offered more information in the inference of the roles. For this purpose, we proposed the conduction of two ablation studies that offered this information. An ablation study, in this field, consists of eliminating features of a dataset or the algorithm to see which of them are the ones that contribute the most to carrying out the proposed task. In the first ablation study, we created graphs using only one node's centrality as a feature for the classification algorithm. The classification performance showed which centralities provided the most information in the classification and the differences between organisations, as seen in Table 5.5.

The second phase of the ablation study analyses the information that each centrality offers to classify each role. To obtain the results of each role, we change the experimental procedure and classify them as in a binary classification problem in which the classifier should decide if a specific user belongs to each of the roles. Through this experiment, we could verify how each proposed centrality contributes when classifying each role. The results of this ablation study are presented in Table 5.4. Thanks to these studies, conclusions can be drawn about the centralities concerning the organisations' roles and structure.

Finally, we evaluate our approach in unconventional scenarios, such as generalisation and semi-supervised scenarios. In the first evaluation, we present our approach to a generalisation problem. We trained our model with two of the three proposed graphs for this experimental process and tested on the third and unseen one. Table 5.6 presents the results obtained from this evaluation. This evaluation aims to understand whether the selected organisations have similar structures and whether our algorithm can understand them and

translate them to previously unseen data, thus, presenting a dangerous attack vector on Twitter.

In the semi-supervised scenario, the training set is reduced to analyse whether our approach can achieve good results when the annotated data is insufficient. To this end, we evaluated the model using training sets representing 2%, 5% and 10% of the whole dataset. The results of these evaluations and the comparison with the state-of-the-art can be seen in Figure 5.2.

To ensure the reliability of our evaluation, we followed the Stratified 10-Folds cross-validation method and calculated the average for the results of all the folds. In each fold, the train and test sets contain 90% and 10% of the dataset nodes, respectively, and they follow the same distribution as the whole dataset. Then, we used the set of centralities explained in Section 3.1 to test the proposed approaches. The centralities were calculated using *NetworkX*(Hagberg et al., 2008) package for Python. All the models, except the GCN, were implemented with the *Scikit-learn* (Pedregosa et al., 2011) package developed for Python. Most of the models use their default parameters except the following ones; in RF, we set the *n_estimators* = 5; in DTC, we set the *min_samples_leaf* = 8 and in kNN, we set the *n_neighbors* to 3. As explained in Section 5.1 we calculate the macro *F-score* metric in the test set.

5.3.2 Dataset

For this evaluation, we collected data from Twitter accounts belonging to three companies' employees. The process of data collection is divided into two phases. The first is to gather users that may be employees of the desired companies; the second is to consult if the users retrieved are actual company employees and check their roles by visiting their LinkedIn profiles. Figure 5.1 illustrates a chart summarising the employed methodology.

The data we used for this study were retrieved using the library and algorithm presented in Section 5.2. Through this method, represented in green in Figure 5.1, we gathered data from three different organisations present on Twitter. A summary of the number of employees per organisation and the ac-

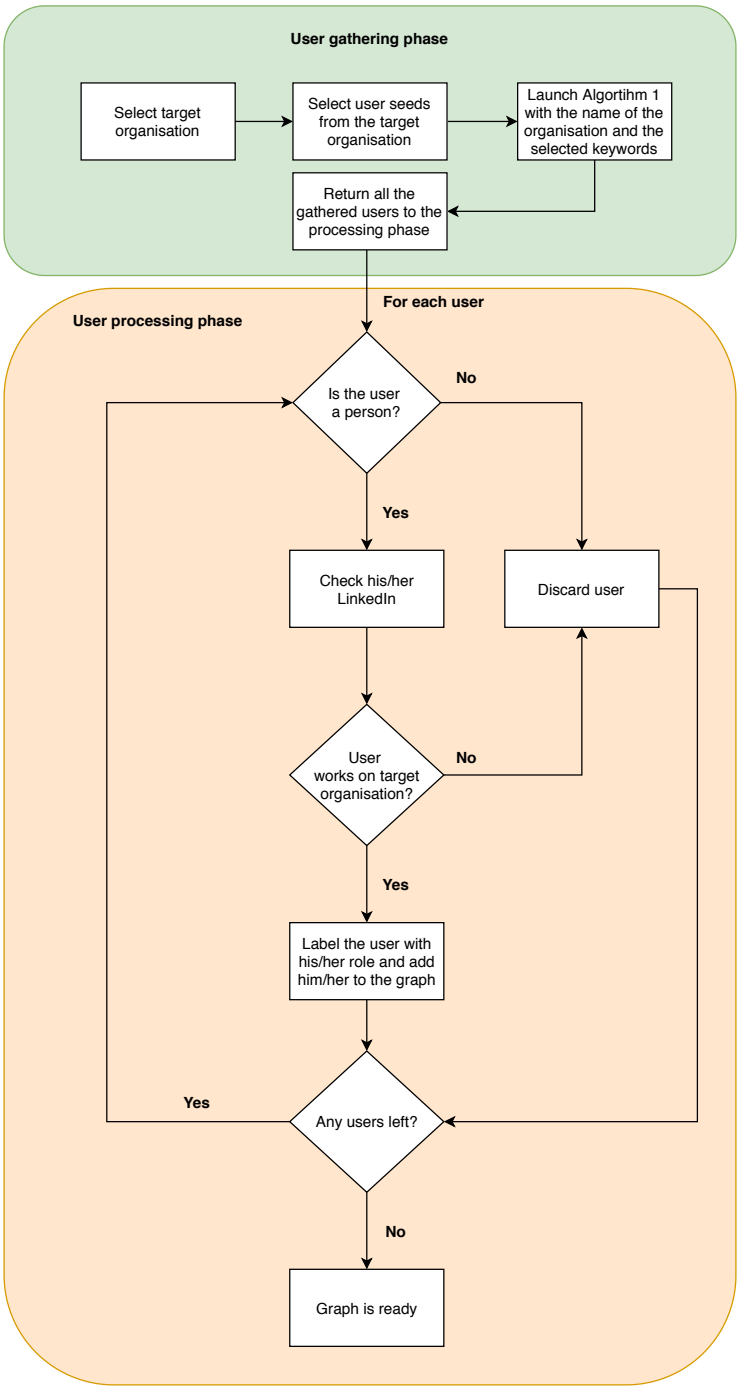

Figure 5.1: Methodology employed for the data collection phase.

Organisation	Nº of accounts	Nº of employees	Nº of non-related accounts	Accuracy	Nodes	Edges
A	2,988	2,706	204 + 78	90.5%	2522	36101
B	1,143	807	163 +173	70.6%	762	3201
C	834	640	78 + 116	76.7%	585	5482

Table 5.1: Summary of the gathered data. Non-related accounts contain corporate accounts and other accounts. The nodes are less than the Nº of accounts because some profiles were discarded as we were unable to decide if they belonged to the organisation.

curacy achieved in each can be seen in Table 5.1. These are the characteristics of the selected organisations and the details of the gathered data:

- *Organisation A* is a multinational professional services firm with more than 200,000 employees, according to its website. The crawler gathered 2,988 different Twitter accounts in a week: 2,706 employees' accounts, 204 corporate accounts and 78 accounts that were somehow related to the company but were not employees.

- *Organisation B* is a multinational engineering company that employs more than 350,000 people worldwide. In this case, 1,143 accounts were gathered; 807 belonged to the organisation's employees, 163 were corporate accounts, and 173 were unrelated to the organisation.

- *Organisation C* is a multinational corporation that provides communication technology and services worldwide and, according to its website, employs 116,000 people. 834 accounts were gathered: 603 from employees, 78 corporate accounts and 116 accounts that were not company employees.

After the crawling finished, we conducted a preprocessing, cleaning and labelling process of the gathered data. We tagged every user in the dataset with its role in the organisation following the methodology represented in orange in Figure 5.1. Users were classified in three roles depending on their position in the company from high to low in the hierarchy: executives, managers and employees. This process was done using public data from LinkedIn. During this process, we had to discard 284 profiles because we could not decide if they

Organisation	Employees	Managers	Executives
A	1787 (70.85%)	698 (27.68%)	37 (1.47%)
B	386 (50.65%)	319 (41.86%)	57 (7.49%)
C	408 (69.74%)	165 (28.21%)	12 (2.05%)

Table 5.2: Distribution of roles per organisation.

belonged to the organisation since the information was not public or was very ambiguous. Table 5.2 show the distribution per class in each organisation.

Finally, after preprocessing and cleaning the collected data, we transformed the data from each organisation into a directed graph. Unlike Facebook, Twitter is not a reciprocal social network, so users do not have to follow each other in pairs, but one user can follow another without the latter following the former. In the previous work presented by (Fire and Puzis, 2016) they employed undirected graphs presenting a difference with our approach. Furthermore, the graphs extracted from Twitter represent three different organisations with different sizes that support the testing of hypotheses in different situations and thus make the results more representative of different realities.

Table 5.1 the algorithm achieved high accuracy rates in the crawls conducted to gather users from the three organisations. The accuracy was calculated by dividing the number of employees gathered by the number of accounts gathered. As we can see in the table, the accuracy for the first organisation (O1) is higher than the average. The main reason is that we used more seeds for the initial step of the crawling algorithm. Also, the accuracy may be limited by employees' descriptions in their online profiles, which is an external limitation. Furthermore, we also gathered corporate accounts that were not taken into account previously by Fire and Puzis (2016).

5.3.3 Discussion

The results obtained in this evaluation show that our approach is more effective than the one proposed in the state-of-the-art for classifying the position of an individual in the hierarchy of an organisation present on Twitter. The performance of many supervised machine learning classifiers presented in Table

	Dataset	KNN (1)	KNN (3)	KNN (10)	NB	DTC	RF	MLP	GCN
Fire	Organisation A	65.78	68.06	68.89	67.87	67.11	69.37	68.28	58.77
	Organisation B	54.52	57.34	56.86	50.70	56.9	59.13	59.03	58.06
	Organisation C	68.52	76.48	77.67	71.86	75.31	77.09	75.25	72.82
	AVERAGE	62.94	**67.29**	67.81	63.48	66.44	68.53	**67.52**	63.22
Ours	Organisation A	67.18	69.65	71.14	69.06	72.45	76.14	69.93	67.70
	Organisation B	52.37	55.63	60.16	52.08	59.12	59.99	55.13	59.51
	Organisation C	73.63	73.37	74.39	71.60	75.57	75.86	72.80	75.37
	AVERAGE	**64.39**	66.22	**68.56**	**64.25**	**69.05**	**70.67**	65.95	**67.53**

Table 5.3: F1 results per algorithm and organisation with our proposed method and the one presented by Fire and Puzis (Fire and Puzis, 2016).

Dataset	Class	Deg.	Indeg.	Outdeg.	Clos.	Betw.	Eigen.	Auth.	Hubs.	Page.
Organisation A	employee	82.37	81.90	**84.10**	76.99	78.54	75.24	78.71	75.56	79.62
	manager	25.82	14.62	34.66	36.24	30.29	34.56	36.47	30.87	**39.50**
	executive	9.05	0	18.13	21.90	2.50	5.33	17.66	0	**25.53**
Organisation B	employee	63.75	65.41	66.58	67.91	67.70	**70.03**	65.14	63.99	65.49
	manager	44.63	**51.81**	46.61	49.41	44.36	49.48	45.81	40.95	44.35
	executive	14.60	0	6.94	**27.49**	16.19	15.33	13.06	8.78	15.69
Organisation C	employee	85.46	85.60	86.39	84.33	83.84	83.83	**86.58**	84.25	84.51
	manager	49.04	50.63	52.80	55.38	49.26	54.37	**58.42**	49.60	49.67
	executive	16.67	0	6.67	**26.67**	6.67	11.67	2.00	0	13.33

Table 5.4: F1 results per organisation and class for each centrality with *Random Forest* machine learning algorithm.

5.3 affirms the possibility of inferring the desired information, thus confirming the RQ 2.a.1. As the table depicts, our proposed set of centralities and the directional graph outperform the work done by Fire and Puzis (2016). Analysing the results shown in Table 5.3 we can ensure that most of the algorithms tested in it with the directional graphs and the set of centralities in our approach perform better than Fire and Puzis proposal, being RF the best one.

Tables 5.4 and 5.5 show the performance obtained in the ablation studies divided per class and dataset separately. We also carried out combinatorial evaluations testing all the centrality combinations and concluded that none of them makes a significant difference from the presented one; a visual representation of the results of these combinatorial experiments are presented in Appendix A.

In the first ablation study, we analysed which centrality offers more information to classify users in their roles. As the Table 5.5 depicts, the *outdegree* centrality slightly outperforms the results obtained by the rest of them. Delving deeper in the results obtained we can compare the performance of two groups of centralities that express concepts of the same scope: *indegree* vs *outdegree* and *closeness* vs *betweenness*. Analysing the degree centralities, we see that the *outdegree* centrality is more important than the *indegree* centrality. We argue that the users that a user follows (followees), represented by *outdegree* centrality, offer more information about their role than the people they are followed by (followers), represented by *indegree* centrality.

On the other hand, *closeness* and *betweenness* measure the distance from the origin node to every other node in the graph and how a node acts as a bridge among the others, respectively. In this case, we can see how having

Dataset	**Deg.**	**Indeg.**	**Outdeg.**	**Clos.**	**Betw.**	**Eigen.**	**Auth.**	**Hubs**	**Page.**
Organisation A	65.57	62.24	69.71	65.39	64.07	62.71	66.19	62.74	67.23
Organisation B	51.54	54.50	54.12	56.59	53.92	57.68	52.47	50.15	52.43
Organisation C	75.35	73.90	76.19	74.89	72.70	74.23	77.08	72.94	73.41
AVERAGE	64.16	63.55	**66.67**	65.62	63.56	64.87	65.25	61.94	64.36

Table 5.5: F1 results of the ablation study conducted to each of the centralities for each of the organisations.

short paths to every other node on the graph is more important to classify a user than the information on how they act as bridges between nodes. We suspect this phenomenon occurs because people with strong influence in the organisation and, hence, executives have the greatest *closeness* centrality in the graph.

Table 5.4 shows the performance of the RF algorithm using centralities individually to classify all the roles separately as a binary problem. Analysing the centralities by categories, we can conclude, by taking the highest performance, that to classify the roles for the employees and managers, Katz-based centralities are the best. Instead, for the executives, the global centralities are the ones that obtained the best performance and the local-based ones the lowest performance. We believe that this is because executives do not have collective behaviour when followed by other users; therefore, we argue that we need to use the global centralities to infer information from this user.

Analysing the centralities individually, we can observe that they behave differently for each role. Although we stated before that the global-based centralities are the ones that achieve better performance when classifying executives, the betweenness centrality is one of the worst for this task; instead, it is one of the best for classifying the employees. We believe this may be because the employees are identified with a low betweenness centrality, and the executives' ones vary. Moreover, as executives used to be connected with more people in the organisation, they usually have high closeness centrality, which is why it is the best for this role.

In addition, the most notable result is that the *indegree* centrality achieves a 0 F-score for every dataset when classifying executives. As we stated before, we assume that these users do not have a collective behaviour when followed by users, which is why this centrality may vary between them. Finally, we can also observe that *hubs* centrality does not work on executive users. We consider that this could happen because this centrality represents the node's capability to be connected with authority nodes that, in this case, and depending on the organisation, could be themselves.

Finally, we evaluate our proposal in generalisation and semi-supervised scenarios compared with the approach proposed by Fire and Puzis (2016).

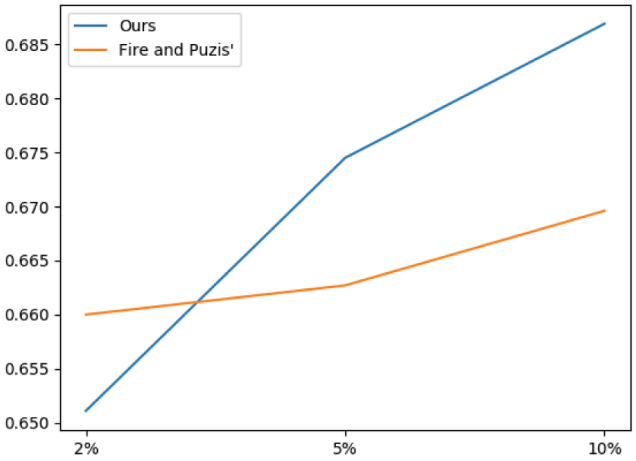

Figure 5.2: Results for the semi-supervised scenario. The percentage of the whole dataset used for training the model is shown on the X axis and the F1 score is represented on the Y axis.

The results presented in Table 5.6 show that our approach can be generalised and, therefore, could be used to classify more organisations without having to be continuously trained, which would allow us to drop the ground truth search from the pipeline. The fact that our model can generalise to these three datasets suggests that if we continue to add data from new organisations, its generalisation performance will increase. In this way, we can extract data from an organisation's users and feed them to the model to know their roles in their organisations without looking for the actual label on another social network.

Dataset	Ours	Fire and Puzis
Organisation A	67.47	**67.97**
Organisation B	**57.26**	47.28
Organisation C	**74.26**	62.57
AVERAGE	**66.33**	59.27

Table 5.6: F1 results for generalisation experiments. The organisation represents the one used as test set.

In the semi-supervised scenario, which results are presented in figure 5.2, our proposed approach performs slightly better than the previously proposed approach except in the first evaluation with 2% of the data for the training phase. The important part to consider here is that our proposed set of centralities and the directed graph work better with larger quantities of data. Moreover, our approach rapidly improves its performance as data increases due to the information that the directed graph and the centralities provide.

Regarding the applicability of this method, we believe that the results obtained in this evaluation process demonstrate the ease with which sensitive data can be extracted from people or organisations on Twitter. Although our analysis is not intended to cause any harm to organisations on Twitter, this method can be applied by malicious users to conduct an attack directed at an organisation by looking for members with higher positions in the hierarchy. Furthermore, such malicious users can also extract information from the organisation and build a more detailed profile to impersonate one of its members. However, this method can also be used by organisations to detect which of their members are influential or to discover new relationships unknown to the organisations and thus create new teams within the company.

With the successful evaluation and the positive answer to both research questions posed for this study, we can ensure the correct fulfilment of the second objective of the dissertation. Through the evaluation, we have demonstrated that the directed graph with our proposed set of centralities can infer the roles of Twitter users in their organisations and perform better than the current state-of-the-art for this task. Furthermore, we also contribute with a theoretical analysis that can be applied in relation-based inference approaches.

5.3.4 Ethical considerations

During the explanation of this analysis in Chapter 3 and throughout this section, we have mentioned numerous times the inference of private information. However, as Twitter states in its terms of use[1], this type of inference is forbidden as it is seen as an attack on user's privacy, as we have discussed previously.

[1] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

However, as we will explain below, we did not infer private information in any study conducted in this thesis.

In this thesis, we propose the study of the inference of private information as a proxy to evaluate the risk that malicious users can exploit the social network to infer information about these characteristics and thus generate attack vectors. Therefore, to follow the terms of use of the social network, we employed public information extracted from other sources (LinkedIn) for the evaluations proposed for the developed methods, and thus, no private information is inferred.

Another potential risk is that malicious users could use the methods and knowledge generated in this analysis to exploit social networks beyond what is currently being done. However, as we stated before, the purpose of this thesis is merely to provide mechanisms to prevent this from happening and to contribute to the resolution of this problem.

5.4 **Evaluation of Organisational Specific Language Analysis**

In this section, we propose that we can differentiate the specific languages that members of organisations use on Twitter and, therefore, the existence of such languages. To validate the second part of the second objective of the dissertation, we posed two research questions (RQ 2.b.1 and RQ 2.b.2). This section will introduce the experimental setup proposed to answer both questions, the dataset used, and a discussion about the obtained results.

5.4.1 **Experimental setup**

To evaluate the objective that we can distinguish between organisation-specific languages on Twitter, we have divided the experimental process into two tasks that answer both research questions posed for this objective: tweet-level (RQ 2.b.1) and user-level (RQ 2.b.2) classification. Although both tasks aim to classify tweets using the model presented above, each partly solves the proposed hypothesis. As previously said, this division responds to the need for

both to be correctly resolved to claim that organisations have a specific language. Thus, on the one hand, the proper classification of the tweet-level task allows us to identify that the languages among the five selected organisations are distinct. However, on the other hand, good results in the user-level task show us that the specific language is shared among all the organisation members.

The tweet-level evaluation aims to analyse if the languages from the tweets of the different companies are distinguishable by the employed models. Therefore, we balanced the total tweets from each organisation, considering the one with the least amount of tweets (390,985 tweets per organisation). Therefore, classifying the tweets with better results than the baseline will mean the existence of a common language among the members of the same organisation.

Furthermore, the user-level evaluation aims to clarify whether the language detected by the employed models is independent of users and corresponds to the language of a specific organisation. For this, we grouped the tweets produced by each user to avoid having tweets of the same user in the training and testing subsets. To balance the tweets per user and the users per organisation, we took 100 tweets from each user and selected the minimum number of users specified by the organisation with fewer users.

We also conducted an ablation study on the tweets. This study helped us analyse the importance of each element in the classification tasks and demonstrate that the organisation's specific language is independent of these elements. Tweets are composed of many elements that help users interact or add more information to the tweet. The elements analysed in this study will be:

- Mention: mentions are used to name other users from Twitter using their usernames (@username). In our analysis, members of the same organisation will usually mention similar users, adding them to the specific language of the organisation.

- Hashtag: hashtags are used to index keywords or topics on Twitter using the key character # (#hashtag). Members of the same organisations will probably use the same hashtags to talk about company topics introducing them to the specific language of the organisation.

- URL: URLs are used to link to external pages. Organisations could use this element to link to similar pages helping to create the specific language of the organisation. The problem regarding URLs in Twitter is that the OSN shortens them, so the name is changed.

To evaluate the employed models, explained in Section 3.2, we split the datasets into three subsets in both tasks. The training subset represents 70% of the whole data, the validation subset represents 10%, and the testing subset represents the remaining 20%. The splits are done randomly for the tweet-level evaluation, considering all available tweets. Instead, in the user-level evaluation, each split is done by each user, so all the tweets produced by one user are in one and only one subset.

We divided our datasets into three subsets because we employed early stopping to stop the training when it starts to over-fit. After each training iteration, the model is tested with the validation set, obtaining a loss metric and saving the model's actual state. Each iteration that loss is compared with the previous one, and when it increases instead of decreasing a limited amount of times (2 in our evaluations), the training phase is stopped. Then the test subset is used to test the model's performance as it has remained unseen until this phase.

We also conducted an ablation study on the tweets. This study assisted us in determining the value of each element in the categorisation tasks and demonstrating that the organisation's unique language is unaffected by them. Tweets contain various components that allow people to interact with the tweet or provide additional information. The elements analysed in this study are (i) mention, (ii) hashtags and (iii) URLs, as explained before.

After, we created a dataset for each combination of none, one or two elements for evaluating each element. Then, we followed the same methodology explained in this section to calculate the model's performance for each dataset. The results obtained for the ablation study can be seen in tables 5.9a and 5.9b and will be discussed in the next section.

Furthermore, we realised that several tweets in our dataset might contain the name of the organisation and the labels in which the model must classify

the tweets. To ensure that the model does not learn the names of the organisations to classify the tweets, we replaced them with a unique token. The results of these ablation studies are shown in Tables 5.9a and 5.9b.

As explained in Section 5.1, we computed the macro F1-score, precision, recall and accuracy in our experimental process. In the Tables 5.8, 5.9a and 5.9b we report the performance for all the evaluation of this study. All the experiments ran in an NVIDIA QUADRO RTX 8000.

5.4.2 Dataset

To evaluate the existence of organisational-specific languages on Twitter, we employed data from Twitter users that identify as members of five selected organisations to classify their tweets or their profiles in one of the five organisations. We chose organisations from different fields but with similarities to see if ranking could be done between organisations with potentially similar languages. In addition, we looked for organisations with enough members on Twitter to have enough information for the proposed models. These are the details of the five selected organisations:

- Organisation A is an NGO focused on human rights.
- Organisation B is a multinational aerospace corporation.
- Organisation C is a multinational professional services network.
- Organisation D is a political party.
- Organisation E is a multinational technology company.

For the data collection of the selected organisations, we employed the algorithm explained in Section 5.2. For each organisation, we introduced 20 manually identified seed users and ran the algorithm for two weeks with the names of the organisations as keywords. The number of users and tweets collected for each organisation can be seen in Table 5.7. After collecting the users, we gathered 3200 tweets for each user as Twitter API restricts.

Out of the tweets collected for each user, we filtered English-language tweets for consistency and removed the emojis from them. Many works have improved the performance of text classifying models by changing the emojis present in the text by their contextual meaning (Singh et al., 2019; Chen et al., 2019). Nevertheless, we decided not to change the emojis for their meaning to avoid modifying the original tweets. Furthermore, some emojis may not mean the same in every context, for example, when using irony, and thus, the automatic change of these emojis may alter the meaning of the original tweet. We also removed retweets as they do not represent content written by the author and very similar tweets (e.g. differing only in a hashtag or a URL) due to their redundant content. We changed all the mentions of the organisation names by a tag *($ORG)* for the remaining tweets, intending to avoid direct mentions of the ground truth label. Table 5.7 shows the number of resulting tweets per organisation.

5.4.3 Discussion

After running the evaluation process and analysing the results, we can ensure that we found the answer to the key research questions 2.b.1 and 2.b.2 proposed in Section 3.2. As we can see in Table 5.8, the BERT model achieved more than 75 points in macro F1-score in both classification processes and achieved substantially better results than the Random Forest Classifier, which does not use the language features. Based on the accuracy and the F1 score obtained, we can assure that most users use a language similar to the one used

Organisation	Users	Tweets	Tweets preprocessed
A	2,298	1,992,434	885,975
B	684	620,836	390,985
C	3,591	2,256,832	1,384,715
D	7,374	15,640,422	6,270,414
E	11,099	14,693,852	6,870,413

Table 5.7: Number of users and tweets per organisations. The last column on the right indicates the number of tweets left per organisation after applying the transformations explained in Section 5.4.2.

	Tweet-level		*User-level*	
Model	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**
Random Forest	33,96	34,35%	53,06	56,87%
Bert	**75,66**	**76,17%**	**87,95**	**87,95%**
Multi-CNN	69,83	69,72%	70,74	71,29%

Table 5.8: Results of the tweet-level and user-level experiments with unprocessed tweets.

by their peers than the one used by the rest. We also wanted to analyse if the classifier relies on mentions, hashtags or URLs to classify the tweets.

To analyse the classification between the different organisations, we have extracted a confusion matrix from an iteration at tweet level classification (Figure 5.3). We can see that E is the organisation in which most errors are centralised when predicting the others. We hypothesise that this may be because E is the largest organisation of all, and it may be that its users usually variate in their language or talk about very diverse topics that attract the classification of users from other organisations. On the other hand, organisa-

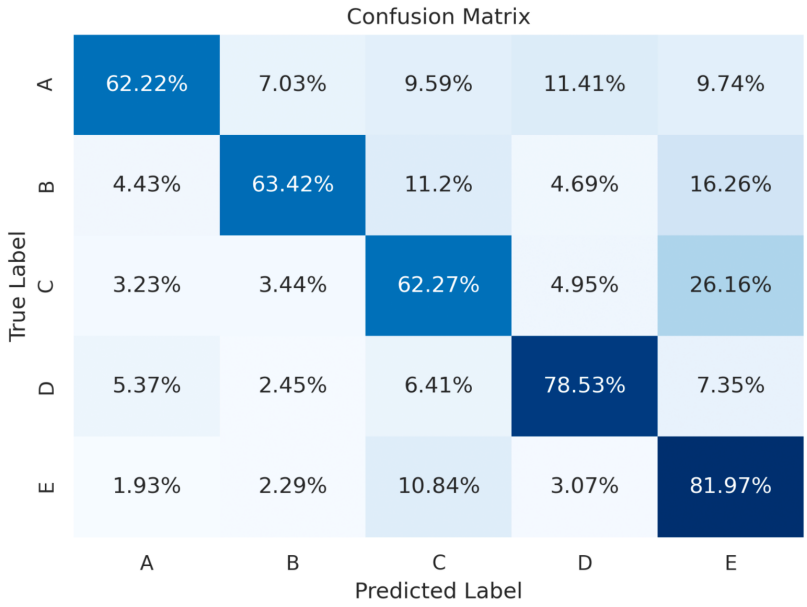

Figure 5.3: Confusion matrix of tweet-level classification with the Multi-CNN model.

tion A is the one that accumulates more errors, so it could mean that their language is not as specific as we thought or that in the tweets selected for the test part, the specific language was not as used as in the train tweets, nevertheless, the results are promising.

The experimentation was divided into two tasks to answer the key research questions proposed and ensure that the language used was common to all the organisation members. First, the tweet-level detection task allows us to analyse the tweets individually and detect a language common to all, thus answering RQ 2.b.1. However, with this approach, the model may focus on specific user styles and forget about the language's style as a whole. To alleviate this problem, the task of user classification was posed. In this second task, the tweets are grouped by user, so the model analyses them together, looking for similarities between users and a common language between them, answering RQ 2.b.2.

Analysing the ablation study conducted on the tweets, which results are presented in Tables 5.9a and 5.9b, we can see how the different combinations of tweet elements affect the classification. Both tweet-level and user-level classifications follow the same importance scale for the tweets' elements with subtle differences. Comparing the model's performance when using only one element in the tweets, we can see that mention is the most important for the classification. This element is used to cite accounts or follow conversations with other users, thus creating a higher homophily bias. In our opinion, mentions contribute significantly to the classification because users of the same organisation talk to each other or quote the same accounts when tweeting. URLs and hashtags are also helping the model achieve better results, but as we can see in the presented tables, they need to be combined with others to improve the performance. Hashtags are used to attach tweets to conversations or topics, and the words used in the hashtags may help the classifier detect words in the specific language of an organisation. URLs are used to link external webs with the possibility that several members of the same organisation link the same web page.

Regarding the ablation studies conducted with two elements on the tweets, it can be seen that the hashtags with the mentions perform better than the

Elements used	*Bert*		*Multi-CNN*		*Random Forest*	
	F1-Score	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**
None - No orgs names	49,48	50,38%	**57,38**	**57,45%**	33,34	33,83%
None with orgs names	**62,37**	**63,44%**	57,75	57,91%	33,61	34,03%
Hashtag	60,61	61,39%	**62,27**	**62,30%**	34,22	34,55%
Mention	**72,35**	**73,20%**	67,01	66,97%	34,37	34,73%
Url	**64,67**	**65,70%**	57,24	57,38%	33,26	33,64%
Hashtag-Mention	**74,65**	**75,37%**	70,87	70,82%	34,71	35,06%
Hashtag-Url	**65,07**	**65,79%**	60,87	60,92%	33,40	33,78%
Mention-Url	**72,80**	**73,67%**	66,22	66,08%	33,61	34,03%
All	**75,66**	**76,17%**	69,83	69,72%	33,96	34,35%

(a) Tweet-level evaluation.

Elements used	*Bert*		*Multi-CNN*		*Random Forest*	
	F1-Score	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**
None - No orgs names	**80,93**	**80,93%**	63,51	65,88%	45,63	51,37%
None with orgs names	**84,66**	**84,66%**	67,27	68,71%	49,87	54,95%
Hashtag	**83,29**	**83,29%**	66,24	67,76%	49,48	54,40%
Mention	**86,85**	**86,85%**	69,71	70,59%	48,88	54,50%
Url	**84,11**	**84,11%**	63,52	65,41%	45,63	51,37%
Hashtag-Mention	**87,12**	**87,12%**	62,18	62,59%	50,65	55,22%
Hashtag-Url	**86,58**	**86,58%**	64,26	64,71%	48,87	54,12%
Mention-Url	**85,75**	**85,75%**	66,42	67,53%	47,47	52,47%
All	**87,95**	**87,95%**	70,74	71,29%	53,06	56,87%

(b) User-level evaluation.

Table 5.9: Results of the evaluations conducted in the ablation study performed on the elements of the tweets in (a) tweet-level and (b) user-level evaluations.

other combinations in both studies. This is because this combination involves two elements that may represent keywords for the model. For example, if a user is always mentioned by members of the same organisation or a hashtag is continuously used by these members, those words would represent a part of the organisation's specific language.

The ablation study supports the idea that the language of the organisations is not only composed of tweet elements. Although, using the elements for the classification enhances the model's performance. The results obtained when replacing all of them, including the organisations' names, demonstrate that the specific language of the organisations is inherent to the elements from Twitter and can be distinguished without them.

Going back to the proposed hypothesis that organisations may have specific languages that their members use on Twitter, we could say that our research strongly supports it by classifying tweets and users under different conditions with good results. This finding may have a theoretical impact on future research related to Twitter attack prevention by generating an attention focus on the language of communities and organisations on Twitter in addition to the language of individuals. Furthermore, through the published dataset, we aim to contribute to implementing or refining systems that aim to detect malicious users before they attack organisations on social networks.

5.5 Evaluation of the Malicious User Foreseeing System

In this section, we propose that malicious users may be preventively detected using a language agnostic model that leverages interactions between them to foresee and detect their behaviour. To validate the fifth objective of the dissertation, we posed two research questions (RQ 5.1 and RQ 5.2). This section will introduce the experimental setup proposed to answer both questions, the dataset used, and a discussion about the obtained results.

5.5.1 Experimental setup

To evaluate the proposed approach presented in Section 4.3, we conducted our experimental process against TGN, the state-of-the-art model in node classification tasks in dynamic graphs introduced below. We will also describe the model selection study that supports the choice of models used and demonstrates their performance.

TGN (Rossi et al., 2020) is the current state-of-the-art for node classification and link prediction for dynamic graphs. The authors propose a model based on combining memory modules to store long-term information and a graph-based embedding module to generate up-to-date node embeddings. One of the major differences with the other models is the use of batches for training, with which the authors demonstrate that they can train the model faster than their competitors. One of the problems that batches can have is that the node representations are not updated until the batch is finished, so the model uses the old node representations. To alleviate this, the authors propose a message and memory aggregation system that considers all interactions in the same batch before generating the node representations. However, we argue that when aggregating batch messages, information about user actions' timing may be lost.

During the development of our model, we performed a model selection study by altering both the embedding foresee model (EFM) and the embedding classification model (ECM) to see which of the proposed algorithms performed best. For both studies, we evaluated the model several times with 10% foresee size and moving t_N up to 50% of the dataset. We ran the evaluation only up to 50% of the dataset interaction to select the models that classified users better without using large percentages of the data in the training sets. Furthermore, one of the proposed research questions (RQ 5.2) refers to the speed with which these users can be detected while they are carrying out the attack, so we seek to use the algorithm that needs the minor data to classify them, as this will be the one that does it the fastest.

For the study conducted to the ECM we used the following classification algorithms: *Support Vector Classifier (SVC)* (Cortes and Vapnik, 1995), *Mul-*

tilayer Perceptron (MLP) (Glorot and Bengio, 2010), *K-Nearest Neighbour (KNN)* (Altman, 1992) and *Random Forest (RF)* (Breiman, 2001). All the algorithms were implemented using the Sckit-learn library with the default hyperparameters. The results of this study are presented in Table 5.13.

After selecting which algorithm was better for the embedding classification model, we evaluated the performance of the embedding foresee model. For this purpose, we modified the MLP originally proposed to classify the embeddings in TGN with the best algorithm selected in the previous analysis. The results obtained by TGN are compared with our approach in Table 5.14.

Subsequently, we validated our approach against TGN in the foreseeing classification task. By validating our approach against the state-of-the-art in the most similar task, we answer the RQ 5.1 proposed in Section 5.1 as the results will answer if it is possible to detect malicious users by their interactions preventively.

For this purpose, we have conducted a user classification evaluation series using the three datasets explained in Section 4.2.2 and different foresee sizes of actions: 10%, 30% and 50% of the total interactions of the dataset. For each iteration, we fixed the foreseeing percentage for the evaluation set and increased the t_N 10% until we reached the final timestamp of the dataset. For example, the Figure 4.3.a will represent the first evaluation experiment of the 10% foreseeing evaluation series and Figure 4.3.b the fourth evaluation of the 30% foreseeing series.

Our approach and TGN models use their default embedding classification models. In TGN, the algorithm proposed by the authors is a custom Multilayer Perceptron (Glorot and Bengio, 2010) classifier. However, we have modified how this MLP is trained and tested to suit our experimental process. As explained in section 4.3, the classifier is trained after the projection of embeddings and not during to avoid data leaking. The train and test split used for the embedding classification are the same for both models.

The following hyperparameters were used for all the proposed evaluations: embedding size of 128, a learning rate of $1e^{-3}$ and a weight decay of $1e^{-5}$. The hyperparameters used for TGN are the default proposed in (Rossi et al., 2020). All the evaluations were run on an NVIDIA RTX 8000 graphic card.

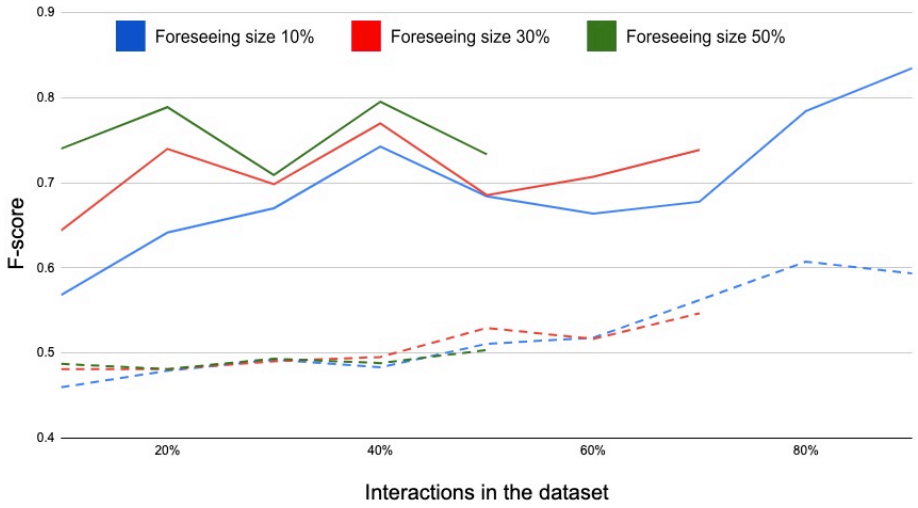

Figure 5.4: Average of the results obtained in the three datasets for both models and with different foresee sizes. The solid lines represent the evaluation of our approach, and the dashed lines the evaluation of TGN.

5.5.2 Discussion

The results demonstrate that our approach outperforms the results obtained by the state-of-the-art model for node classification in dynamic graphs, TGN. Figure 5.4 shows how the average results of our model are more than 20 points above the results obtained by TGN. As we can see in the tables, our model achieves good results in the three datasets proposed for its evaluation. However, in the Russian dataset, our model fails past the middle of the dataset in all the series. Even though its performance is good with a lower foresee size, in the other evaluations, it cannot recover. We argue that this may be due to a sudden change in the behaviour of both malicious and legitimate users that does not correspond to the previous training of the model, causing it to fail in the projection of user embeddings.

In the cited tables, we can see that giving the model a minimum amount of data cannot generate good representations of the users, as the user classification is similar to that of a random classifier. However, adding more interactions to the training set allows our model to achieve classification results above 0.75 F-score before reaching 50% t_N on some datasets. These results

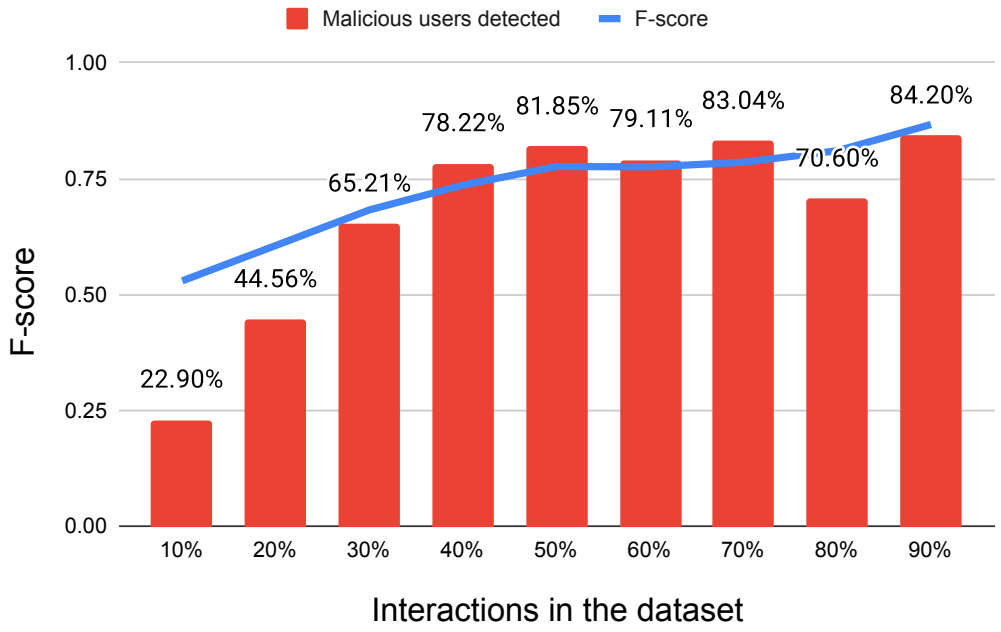

Figure 5.5: Evaluation series conducted with foreseeing size of 10% in the Iran dataset. The red bars represent the percentage of malicious users detected by the model, and the blue line represents the F-score obtained.

validate that the model does not need much training to start projecting representations that the proposed algorithms can successfully classify. We can also see in Figure 5.5 how the model increases its accuracy when detecting malicious users over time, represented as red bars. It can be seen how with only 40% of the dataset interactions, the model can detect more than 75% of the malicious users.

We also analysed the foresee size and its effects on the subsequent classification of embeddings during the experimentation. As explained in the previous section, evaluations were carried out with foreseeing sizes of 10%, 30% and 50% of the total number of interactions in the dataset. In the results reported in Tables 5.10, 5.11 and 5.12, it can be seen how the model achieves similar results with the three different foresee sizes; even as the size increases, we can see that the model achieves better results faster. However, it is also visible that the larger the foresee size, the more susceptible the model is to errors that affect its subsequent classification.

	China		*Iran*		*Russia*	
t_N	**Our approach**	**TGN**	**Our approach**	**TGN**	**Our approach**	**TGN**
10%	**0.6142**	0.4720	**0.5930**	0.4220	**0.6742**	0.4880
20%	**0.6109**	0.4198	**0.6187**	0.4783	**0.6493**	0.5480
30%	**0.6527**	0.4308	**0.6456**	0.5086	**0.6809**	0.5441
40%	**0.6566**	0.4162	**0.701**	0.5398	**0.6244**	0.5025
50%	**0.766**	0.4677	**0.7426**	0.5131	**0.5915**	0.5558
60%	**0.7871**	0.4417	**0.7836**	0.5096	0.5832	**0.6172**
70%	**0.7698**	0.5143	**0.7697**	0.5519	0.5451	**0.6263**
80%	**0.8006**	0.5305	**0.8256**	0.6116	**0.7219**	0.6909
90%	**0.8414**	0.5593	**0.6892**	0.5740	**0.8248**	0.6511

Table 5.10: F-score results for our approach and TGN on the three proposed datasets using a foreseeing size of 10%

	China		*Iran*		*Russia*	
t_N	**Our approach**	**TGN**	**Our approach**	**TGN**	**Our approach**	**TGN**
10%	**0.6103**	0.424	**0.6485**	0.507	**0.7053**	0.518
20%	**0.6269**	0.406	**0.7033**	0.539	**0.6066**	0.509
30%	**0.7183**	0.436	**0.7428**	0.548	**0.6544**	0.494
40%	**0.6544**	0.415	**0.8041**	0.564	**0.6244**	0.519
50%	**0.7869**	0.486	**0.6905**	0.505	0.5999	**0.604**
60%	**0.8427**	0.455	**0.7874**	0.48	0.5719	**0.631**
70%	**0.8882**	0.496	**0.8329**	0.526	0.5137	**0.625**

Table 5.11: F-score results for our approach and TGN on the three proposed datasets using a foreseeing size of 30%

	China		*Iran*		*Russia*	
	Our approach	**TGN**	**Our approach**	**TGN**	**Our approach**	**TGN**
10%	**0.7053**	0.432	**0.7446**	0.542	**0.7243**	0.494
20%	**0.7294**	0.403	**0.807**	0.562	**0.6114**	0.492
30%	**0.7157**	0.457	**0.8052**	0.542	**0.6684**	0.485
40%	**0.6547**	0.45	**0.8082**	0.475	**0.6227**	0.544
50%	**0.8965**	0.468	**0.5279**	0.457	0.5417	**0.597**

Table 5.12: F-score results for our approach and TGN on the three proposed datasets using a foreseeing size of 50%

During the development of our model, we also performed a model selection study by altering both EFM and ECM to see which of the proposed algorithms performed best. For both studies, we conducted evaluations with 10% foresee size and moving t_N up to 50% of the dataset. We ran the evaluations only up to 50% of the dataset interaction to select the models that classified users better without using large percentages of the data in the training sets.

For the study conducted to the embedding classification model we used the following classification algorithms: *Support Vector Classifier (SVC)* (Cortes and Vapnik, 1995), *Multilayer Perceptron (MLP)* (Glorot and Bengio, 2010), *K-Nearest Neighbour (KNN)* (Altman, 1992) and *Random Forest (RF)* (Breiman, 2001). All the algorithms were implemented using the Sckit-learn library with the default hyperparameters. The results of the evaluations conducted in this study are presented in Table 5.13.

After selecting which of the algorithms was better for the embedding classification model, we conducted a study on the embedding foresee model. For this purpose, we modified the MLP originally proposed to classify the embeddings in TGN with the best algorithm selected in the previous evaluations. The results obtained by TGN are compared with our approach in Table 5.14.

The first study shows that the different proposed algorithms can classify the embeddings produced by the forecasting model with good results. However, being aware that one of the most important things for a preventive model is that it is able to classify the users as soon as possible correctly, we will choose the algorithm that classifies the users in the first sections of the dataset with the best result. In this case, the algorithm we used for our approach was Random Forest because of its good results. Although we can see how other algorithms obtain better results in some points of the dataset, the difference with Random Forest, except in the Russian dataset, is not significant enough to change the cloning algorithm.

The results in Table 5.14 show that our approach is superior to TGN, together with the best embedding classifier, in two of the three datasets used. In this way, it can be seen how our EFM manages to represent the users more accurately by updating the representations of actions and users after each interaction. However, our model loses against TNG+RF in the Russia

	China				*Iran*				*Russia*			
t_N	**KNN**	**MLP**	**RF**	**SVC**	**KNN**	**MLP**	**RF**	**SVC**	**KNN**	**MLP**	**RF**	**SVC**
10%	0,5676	0,5621	**0,5930**	0,5534	0,5818	0,6000	0,5294	**0,6311**	0,5832	0,5812	**0,5841**	0,5829
20%	0,6092	0,5876	**0,6298**	0,5859	0,6202	0,6343	0,6057	**0,6455**	0,6631	0,6759	**0,6928**	0,6849
30%	0,6295	0,6400	0,6421	**0,6433**	0,6649	0,6497	**0,6825**	0,6788	0,6662	0,6287	**0,6869**	0,6325
40%	0,6782	0,6446	**0,6963**	0,6438	0,6963	0,7062	**0,7360**	0,7141	**0,8013**	0,7451	0,7990	0,7213
50%	0,6896	0,6673	**0,7192**	0,6757	0,7344	0,7310	**0,7762**	0,7380	**0,6125**	0,5789	0,5737	0,6054

Table 5.13: F-score of the model selection study conducted to the embedding classification model with the three datasets. The foreseeing size was set to 10%, and the t_N was increased 10% after each iteration.

	China		*Iran*		*Russia*	
t_N	**Our approach**	**TGN+RF**	**Our approach**	**TGN+RF**	**Our approach**	**TGN+RF**
10%	0,5930	**0,6193**	**0,5294**	0,5040	0,5841	**0,5879**
20%	**0,6298**	0,5359	**0,6057**	0,5759	0,6928	**0,7022**
30%	**0,6421**	0,5769	**0,6825**	0,6455	0,6869	**0,7191**
40%	**0,6963**	0,6081	**0,7360**	0,7083	0,7990	**0,8128**
50%	**0,7192**	0,6654	**0,7762**	0,7682	0,5737	**0,8575**

Table 5.14: F-score of the model selection study conducted to the embedding foresee model with the three datasets. We changed employed the best performing embedding classification model for both approaches. The foreseeing size was set to 10%, and the t_N was increased 10% after each iteration.

dataset by a few points in the first four evaluations. Furthermore, as we have seen in previous results, it performs much worse when it reaches the middle of the dataset. Nevertheless, we can assure that our model is superior since it obtains better results in two of the three proposed datasets.

5.5.3 Ethical considerations

Even though the creation of this model should be necessary to maintain the security and neutrality of social networks, there is some ethical aspect that should be treated not to impose systems that go against their objectives. The first idea that comes to mind when discussing ethical aspects of social media is freedom of expression. Various institutions, such as Amnesty International, define freedom of expression as the right to express, disseminate, seek, receive, and share information and ideas without fear. Taking that definition at face value, social networks should allow users, regardless of their opinion, to express and disseminate their ideas without limitations or fear of censorship; that is to say, they should apply neutrality to the messages published on their platform. However, in this exercise of tolerance of all ideas, we come up against the paradox of tolerance, which was set out by the Austrian philosopher Karl Popper in 1945. He defined the paradox in his book "Society and its Enemies" as follows: "Unlimited tolerance must lead to the disappearance of tolerance. If we extend unlimited tolerance even to those who are intolerant, if we are not prepared to defend a tolerant society against the

outrages of the intolerant, the result will be the destruction of the tolerant and, like them, of tolerance. (Popper, 1971)" Social networks must therefore ensure that this paradox of tolerance does not occur by limiting certain hate messages by intolerant people.

On the other hand, the need to censor users depending on their behaviour on the social network leaves another ethical debate about deciding which users should be considered malicious. As we have previously discussed in Chapter 4, in most of the datasets created by the scientific community, they were the ones who set the rules on what was considered a bot or not. However, in the case of our model, when trained with the data provided by Twitter through the TTC, the social network sets the criteria. In both cases, there is a common problem: the need for a "jury" to dictate the criteria for defining a user as malicious. For this reason, Twitter considered at the end of 2021 the creation of the Twitter Moderation Research Consortium (TMRC), composed of different entities in academia, civil society, NGOs, and journalism, to study the platform governance issues. To which, as a result of the work carried out through this dissertation, we have been invited.

From the above, we see the need to design and implement models that stop users from polluting the networks and creating hatred towards the ideas of others. Creating these systems with the help of different organisations, as Twitter has proposed, is the solution to avoid biases towards some ideas or forgetting about realities that may occur in the social network itself, of which some people are unaware. These collaborative moderation models and processes are the solution to preserve fair and healthy social networks that are seen as information platforms instead of sources of problems in people's daily life processes and even psychological problems.

However, we are aware that the accuracy of the proposed approach is not perfect and that in a real environment, it may flag users erroneously, damaging their experience on the social network in case they get banned. Therefore, results obtained by applying this model to live Twitter data should be manually analysed to identify possible false positives. In addition, and as we will discuss in future work, our goal for this model is to identify the strategies that malicious users carry out in order to identify groups of these

users and, thus, minimise the damage that these attacks can do when large numbers of users amplify them. Examples of such groups can be found in the data provided by the TTC.

5.6 Summary and Conclusions

This chapter presents the evaluation for the three studies we have carried out during the thesis. The obtained results validate our approaches for the three proposed tasks. Moreover, we also demonstrate, through experimentation, that our approaches perform correctly in different situations, such as semi-supervised or generalisation scenarios. Finally, we also conducted ablation studies in some of the proposed models.

Additionally, we have presented a discussion for each of the evaluations sustained by the results obtained in the evaluation conducted following methods that statistically support them. Furthermore, thanks to the analysis of the results, we have drawn conclusions that will be presented in the next chapter. Finally, although the results obtained have been good, we are aware that they leave us with future work that we have been able to identify and that motivates further research in the area.

Finally, ethical issues arise in the experimentation dedicated to foreseeing malicious users on Twitter. Therefore, we presented a section that addresses those ethical issues, presenting our approach to them. With the ethical issues resolved and due to the growing wave of disinformation and hate in social networks, we could conclude that we see the need for the creation of these models with a clear aim and objectives: the creation of free and independent social networks that are not under the control of third parties or biased by the interests of minorities.

In conclusion, we can state through the evaluation we conducted in this chapter and the results we obtained that our approaches are better than those proposed in the state-of-the-art of the different tasks. Furthermore, we proposed a novel approach to implementing models that better apply to the dynamic nature of Twitter

Hoaxes use weaknesses in human behaviour to ensure they are replicated and distributed. In other words, hoaxes prey on the Human Operating System.

Stewart Kirkpatrick

CHAPTER

6

Conclusions and Future Work

This chapter provides an overview of the major findings and contributions given throughout the dissertation. As a result, in order to complete this dissertation, the objectives listed in Chapter 1 are revisited to assess the degree to which they have been met. A review of the major contributions made by this research study, as well as a list of publications relevant to this PhD dissertation, are also included to demonstrate that the scientific community has validated this research. The chapter concludes with some suggestions for future research on coordinated attacks and malicious user detection on Twitter

The rest of the chapter comprises various sections, starting from Section 6.1, which summarises the work done and its conclusions. Then, in 6.2 provides a list of the main contributions of this dissertation. Section 6.3 explains how the objectives proposed at the beginning of the dissertation have been achieved. Then, in section 6.4, we list all the scientific publications obtained during the development of this dissertation. Subsequently, in Section 6.5, we propose the future work that remains after the completion of this

thesis. Finally, we conclude this chapter in Section 6.6, which includes the final remarks and insights gathered throughout this part of the work.

6.1 Summary of Work and Conclusions

This dissertation begins with an analysis of how open source intelligence can help to infer private information from communities and unveil attack vectors against them. These attack vectors are the security breaches malicious users employ to carry out their attacks more easily against communities on Twitter. An example of this is the psychological bias called homophily, which makes people tend to bond with people with whom they already share interests (race, ethnicity, gender or age, for example). In these previous works, we demonstrated the presence of several attack vectors with the inference of private data or the presence of specific languages for specific groups of people. Therefore, following this work, we have designed and implemented a model capable of preemptively detecting malicious users to reduce the damage their attacks can do.

The first two analyses conducted by leveraging the open source intelligence on Twitter were the following: inference of organisational hierarchy through the relationships between its members on Twitter and an analysis of the existence of an organisation-specific language used by organisational members.

Concerning the work on inferring the hierarchy of organisations through the relationships between their members on Twitter, we based our work on the previous state-of-the-art conducted on Facebook by (Fire and Puzis, 2016). In our work, we presented a connection-based inference method that deduces the hierarchical roles of employees in organisations using their relationships on Twitter. To this end, we used data scraped from Twitter and LinkedIn to create directed graphs that allow us to use different centralities and extract more information about the employees' relationships in the OSN. We used these features to classify each employee into one of the three roles proposed, thus increasing the classification roles from the previously proposed approach by (Fire and Puzis, 2016) and using supervised learning methods such as Random Forest.

Through this work, we present an in-depth analysis of the proposed set of centralities to ascertain how they perform when classifying users into their roles. The proposed validation presented in Section 5.3 demonstrates that for most users, outdegree centrality, therefore, the follows that a user has, is the most crucial information to classify him/her in his/her correct role. Regarding performance in different scenarios, our method has achieved considerably good results when reducing the training set, proving that it can be used with a small quantity of annotated data. Besides, we also tested the performance of our algorithm when testing in an unseen dataset; this will allow us to use it for new organisations without having to retrain it each time.

Regarding the second analysis, research in Section 3.2 proposes that language may represent an attack vector usable by malicious users to conduct impersonation or infiltration attacks against communities on Twitter. For its verification, we gathered tweets from members of five different organisations with a presence on Twitter. Subsequently, we divided the experimentation into two tasks: (i) tweet-level task and (ii) user-level task. These tasks ensure that the classification was made by language and not by other factors, such as the style of a unique person. For the experiments proposed, we employed a Convolutional Neural Network-based model for text classification that obtained almost 75% of accuracy in both tasks.

We also conducted an ablation study to analyse which elements present on a tweet (hashtags, mentions or URLs) offer more information about the organisation its author is a member. Besides confirming that mentions and hashtags are the most crucial elements when classifying tweets, we confirmed that the selected organisations' specific languages might be classified without relying on these elements. Furthermore, the successful results obtained by our algorithm when all the tweet elements and the name of the organisation were changed for tokens show that members of the same organisation share an inherent writing style independent of hashtags, mentions and URLs. The validation presented in Chapter 5 demonstrates that organisations have specific Twitter language used by their members, which may be used as an attack vector.

Finally, we present a novel approach for detecting malicious users preemptively, i.e. before they finish their attacks. The proposed approach, explained in Chapter 4, uses user interactions and features extracted from the URLs and hashtags in their tweets to detect malicious users. The model enables the detection of these users either by detecting that they have performed malicious actions or by spotting that the user is performing early actions that resemble the patterns observed in previously seen malicious users. In addition, we present a different methodology from those used in these models by training the model by following the temporal patterns of the social network; thus, this allows them to discover how the user's actions evolve.

For the validation of our approach, we bring the state-of-the-art for node classification in dynamic graphs to the problem of malicious user detection, framed for the first time as a preemptive task. In Section 5.4, we demonstrate the effectiveness of the proposed model for the task by achieving competitive performance on the proposed benchmarks. We have also found that the model is able to project users' embeddings over the longer term without significantly reducing the subsequent classification result. Furthermore, we set up several experiments to test the feasibility of different classification algorithms for the proposed model and chose the one that performed best on the proposed benchmarks. Finally, we have also carried out an ablation study in which we quantify the difference in classification if we remove the features of the tweets when classifying users. This ablation study shows that features do not provide much information to classify users and can be eliminated to create a model based simply on interactions. In this way, the model would be completely language agnostic.

6.2 **Contributions**

A summary of the contributions made during the completion of this thesis are presented in this section:

- *It has been demonstrated that it is possible to infer the roles of members of an organisation through their relationships on Twitter.* This contribution addresses objective 2.a.

- *A dataset containing three graphs composed of users from 3 different organisations united by their relationships on Twitter(Sánchez-Corcuera, 2022c).* This dataset was employed to evaluate the model presented in Section 3.1. This contribution has been achieved following objective 2.a.

- *It has been proven that members of organisations use different languages that may represent specific languages of organisations. Elements on tweets have also been analysed, and the difference in classifying individual tweets and all the tweets from a user.* This contribution has been achieved following objective 2.b.

- *A dataset containing 41,545,828 tweets from members of five different organisations (Sánchez-Corcuera, 2022b).* This dataset was employed to evaluate the model presented in Section 3.2, which was capable of classifying tweets and members of different organisations. This contribution has been achieved following objective 2.b.

- *A software that enhances Twitter's API data collection capabilities was developed and released to the public.* This contribution addresses objective 3.

- *A novel approach that leverages Twitter interactions between users to foresee which will become malicious after a specific set of actions.* This contribution has been achieved following objective 5.

- *A new methodology has been proposed to train and test malicious user detection models in a preventive way. The proposed methodology uses the data respecting the temporality in which they were created.* This contribution has been achieved by addressing objective 4, which was subsequently used for objective 5 and its contributions.

- *A dataset containing 596,221 tweets from legit users related to malicious users from three state-backed operations detected by Twitter and offered publicly in TTC (Sánchez-Corcuera, 2022a).* This dataset has been created to train and test the model proposed in objective 5 of the thesis.

6.3 Hypothesis and objective validation

In the first chapter of this dissertation, the following hypothesis was posed:

> **Hypothesis.** *Using interactions that users make on Twitter, it is possible to create a language agnostic system that foresees malicious users' behaviour and detects them in a preventive way.*

A goal was also created in order to be able to validate this hypothesis, which is also listed below for convenience:

> **Goal.** *Design, implement and validate a language agnostic model capable of foreseeing users' behaviour on Twitter to detect possible malicious users who intend to deploy idea induction attacks.*

O1 *To study the current state of the art on malicious user profiling on OSNs and the automated malicious user detection systems on Twitter.* In Chapter 2, we presented a review of the current state-of-the-art on malicious user profiling and detection systems on Twitter. We also analysed the state-of-the-art in private information inference as these represent an attack vector used by malicious users in their attacks.

O2 *To analyse and leverage the open source intelligence on Twitter to infer information from communities on Twitter.* In chapter 3, we present the analyses carried out to demonstrate the possibility of inferring private data from public data collected from the social network.

- **O2.a** *To design, implement and evaluate a model for inferring the hierarchical roles of users in the organisations from which they are members using information from their Twitter relationships.* In Section 3.1, the design and implementation of the model used to infer the hierarchical structure of the selected organisations is detailed. The validation of this model is presented in the Section 5.3.

O2.b *To design, implement and evaluate a model capable of differentiating organisation-specific languages using Twitter data from their members.* In Section 3.2, the design and implementation of the model used to classify tweets of different organisations is detailed. The validation of the proposed model is presented in Section 5.4.

O3 *To design and implement a time-unrestricted Twitter data capture tool to complement the data provided by the TTC.* In Section 5.2, the implementation of the library designed to capture the data necessary to carry out the thesis is detailed. This section presents the main criticisms of the Twitter API that led to the development of this tool and presents other software created by the community.

O4 *To identify an appropriate evaluation methodology for the malicious user foreseeing model task with its correspondent metrics and perform a quantitative analysis of the results.* Section 5.1 details the metrics proposed for the evaluation of the proposed models. The methodology employed for each model is also presented in the evaluation section of each model.

O5 *To design, implement and evaluate a supervised deep learning model that employs interactions between Twitter users to foresee their behaviour and classify them as malicious users pre-emptively.* Chapter 4 presents the motivation, design and implementation for the creation of this model. Section 5.5 details the evaluation process followed to validate the proposed approach.

6.4 Relevant Publications

The following scientific manuscripts were presented to the scientific community and published in relevant international forums, such as indexed journals and conferences, during the production of this dissertation.

6.4.1 International JCR Journals

An experiment on the inference of private information of users and organisations on Twitter.

- Sánchez-Corcuera, R., Bilbao-Jayo, A., Zulaika, U., & Almeida, A. (2021). Analysing centralities for organisational role inference in online social networks. Engineering Applications of Artificial Intelligence, 99, 104129.

A novel approach to prove the existence of specific languages of organisations on Twitter that malicious users can employ as an attack vector.

- Sánchez-Corcuera, R., Zubiaga, A., & Almeida, A. (2021). Analyzing the Existence of Organization Specific Languages on Twitter. IEEE Access, 9, 111463-111471.

A new technique for link weight prediction that learns from graph structure features and link relationship patterns.

- Zulaika, U., Sánchez-Corcuera, R., Almeida, A., & López-de-Ipiña, D. (2022). LWP-WL: Link weight prediction based on CNNs and the Weisfeiler–Lehman algorithm. Applied Soft Computing, 120, 108657.

6.4.2 International Conferences

A new technique for link weight prediction that learns from graph structure features and link relationship patterns.

- Sánchez-Corcuera, R., Zubiaga, A. & Almeida, A. (2022). Achieving Participatory Smart Cities by Making Social Networks Safer. In 7th International Conference on Smart and Sustainable Technologies (SpliTech 22') (pp. 1-6).

6.4.3 Research grants

As a result of the work carried out in this thesis, the INCEPTION research project has been written and accepted by the Spanish Ministry of Science and Innovation (PID2021-128969OB-I00). The project has received a grant of 101,035€ to continue working in the area of malicious user detection on Twitter and will also have a research grant for a PhD student.

6.4.4 Datasets

- A dataset containig 3 graphs composed of users from 3 organisations united by their relationships on Twitter (Sánchez-Corcuera, 2022c).
- A dataset containing 41,545,828 tweets from members of five different organisations (Sánchez-Corcuera, 2022b).
- A dataset with 596,221 tweets from legit users that complement malicious data extracted from the TTC (Sánchez-Corcuera, 2022a).

6.4.5 Technical Contributions

- A python library called TPL that allows to download and process tweets and prepare them for the models presented in this dissertation[1].
- The source code of the model proposed to detect malicious users preemptively[2].

6.5 Future Work

Inspired by the limitations of the research presented in this dissertation, we have identified the following further research lines:

[1] https://github.com/rubensancor/TPL
[2] https://github.com/rubensancor/Mondrian

- Regarding the leverage of open source intelligence on Twitter to detect organisational specific languages, we detected that it would be interesting to be able to create a language model that would be able to create tweets in the same style as the members of a specific organisation, as surveyed in (Jin et al., 2022). This language model would intend to create a model capable of differentiating between artificial tweets created by the language model and those created by real users. This classification model could be used as a plugin for detecting messages styled by malicious users.
- In this thesis, we worked on detecting malicious users in a preventive way by using a language agnostic model that leverages the information of their interactions on Twitter. However, we believe it would be helpful to focus our efforts on detecting coordinated movements or communities of malicious users that share features. Detecting communities of malicious users or coordinated movements on Twitter would avoid focusing on individuals, which is proving to be a more and more difficult task.
- It would also be interesting to analyse the strategies malicious users use to carry out their attacks. By creating a taxonomy of these strategies and analysing their movements, the model could focus on strategies malicious users from different countries may use and generalise among them. Finally, this taxonomy and analysis of strategies can also help legitimate Twitter users to detect and flag them.
- Finally, it would be helpful for the security of social networks to create a model that detects the point at which malicious users start the attack and become attackers. To create such a model, it would be necessary to have pre-attack data on users that are not reported in the TTC and cannot be obtained by suspending Twitter users.

6.6 Final remarks

This dissertation presents an initial analysis of the related work related to profiling and detecting malicious users. Furthermore, it analyses the state-of-

the-art of open source intelligence exploitation for private information inference and the unveiling of attack vectors against communities. Therefore we demonstrated the ease with which we inferred the hierarchical roles of Twitter users in their organisations and the existence of specific languages that malicious users may use, along with psychological biases such as homophily.

Motivated by this, we finally decided to design and implement a novel language agnostic model capable of preemptively detecting malicious users using their Twitter interactions. We also employed a novel preventive approach for its training and testing that differs from the typical forensic approach proposed in the current state-of-the-art.

Due to the continuous attacks received on Twitter and the lack of application of the current models, we believe that the bot detection strategy used so far should be changed to focus on any type of malicious user regardless of how they are managed and using a preventive approach to reduce the damage the attacks may have on societies. That is why we are committed to using preventive methods that can detect these users before the attack and reduce the damage.

Bibliography

Abreu, J. V. F., Ralha, C. G., and Gondim, J. J. C. (2020). Twitter bot detection with reduced feature set. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

Aichner, T., Grünfelder, M., Maurer, O., and Jegeni, D. (2021). Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Argamon-Engelson, S., Koppel, M., and Avneri, G. (1998). Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4.

Bahri, M., Ashino, R., and Vaillancourt, R. (2013). Convolution theorems for quaternion fourier transform: properties and applications. In *Abstract and Applied Analysis*, volume 2013. Hindawi.

Bangcharoensap, P., Kobayashi, H., Shimizu, N., Yamauchi, S., and Murata, T. (2015). Two step graph-based semi-supervised learning for online auction fraud detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 165–179. Springer.

Bastos, M. T. and Mercea, D. (2019). The brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1):38–54.

Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J. Y., Soulier, L., San-Juan, E., Cappellato, L., and Ferro, N. (2018). Experimental ir meets multilinguality, multimodality, and interaction. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Lecture Notes in Computer Science (LNCS)*, volume 11018, pages 267–285. Springer.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12.

Beskow, D. M. and Carley, K. M. (2018). Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *Conference paper. SBP-BRiMS: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, volume 3, page 3.

Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(7).

Bevensee, E. and Ross, A. R. (2018). The alt-right and global information warfare. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4393–4402. IEEE.

Bilbao-Jayo, A. and Almeida, A. (2018). Political discourse classification in social networks using context sensitive convolutional neural networks. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 76–85.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1):55–71.

Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118.

Bovet, A. and Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., and Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384.

Chaniotakis, E., Antoniou, C., Aifadopoulou, G., and Dimitriou, L. (2017). Inferring activities from social media data. *Transportation research record*, 2666(1):29–37.

Chen, C., Zhang, J., Xie, Y., Xiang, Y., Zhou, W., Hassan, M. M., AlElaiwi, A., and Alrubaian, M. (2015). A performance evaluation of machine learning-based streaming spam tweets detection. *IEEE Transactions on Computational social systems*, 2(3):65–76.

Chen, J., He, J., Cai, L., and Pan, J. (2016). Profiling online social network users via relationships and network characteristics. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE.

Chen, Z., Shen, S., Hu, Z., Lu, X., Mei, Q., and Liu, X. (2019). Emoji-powered representation learning for cross-lingual sentiment classification. In *The World Wide Web Conference*, pages 251–262.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on dependable and secure computing*, 9(6):811–824.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2015). Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2016). Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., and Tesconi, M. (2018). $fake: Evidence of spam and bot activity in stock microblogs on twitter. In *Twelfth international AAAI conference on web and social media*.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., and Tesconi, M. (2019a). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, 13(2):1–27.

Cresci, S., Petrocchi, M., Spognardi, A., and Tognazzi, S. (2019b). Better safe than sorry: an adversarial approach to improve social bot detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 47–56.

Dai, H., Wang, Y., Trivedi, R., and Song, L. (2016). Deep coevolutionary network: Embedding user and item features for recommendation. *arXiv preprint arXiv:1609.03675*.

Das, A., Gollapudi, S., Kıcıman, E., and Varol, O. (2016). Information dissemination in heterogeneous-intent networks. In *Proceedings of the 8th ACM Conference on Web Science*, pages 259–268.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dey, P., Chaterjee, A., and Roy, S. (2019). Influence maximization in online social network using different centrality measures as seed node of information propagation. *Sādhanā*, 44(9):1–13.

Diallo, S. Y., Lynch, C. J., Gore, R., and Padilla, J. J. (2016). Identifying key papers within a journal via network centrality measures. *Scientometrics*, 107(3):1005–1020.

Echeverría, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Stringhini, G., and Zhou, S. (2018). Lobo: Evaluation of generalization deficiencies in twitter bot classifiers. In *Proceedings of the 34th annual computer security applications conference*, pages 137–146.

Enli, G. (2017). Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. *European journal of communication*, 32(1):50–61.

Fani, H., Jiang, E., Bagheri, E., Al-Obeidat, F., Du, W., and Kargar, M. (2019). User community detection via embedding of social network structure and temporal content. *Information Processing & Management*, page 102056.

Feng, S., Tan, Z., Li, R., and Luo, M. (2022a). Heterogeneity-aware twitter bot detection with relational graph transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3977–3985.

Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., et al. (2022b). Twibot-22: Towards graph-based twitter bot detection. *arXiv preprint arXiv:2206.04564*.

Feng, S., Wan, H., Wang, N., Li, J., and Luo, M. (2021a). Twibot-20: A comprehensive twitter bot detection benchmark. *arXiv preprint arXiv:2106.13088*.

Feng, S., Wan, H., Wang, N., and Luo, M. (2021b). Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 236–239.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Fire, M. and Puzis, R. (2016). Organization mining using online social networks. *Networks and Spatial Economics*, 16(2):545–578.

Fourkioti, O., Symeonidis, S., and Arampatzis, A. (2019). Language models and fusion for authorship attribution. *Information Processing & Management*, 56(6).

Freitas, C., Benevenuto, F., Veloso, A., and Ghosh, S. (2016). An empirical study of socialbot infiltration strategies in the twitter social network. *Social Network Analysis and Mining*, 6(23).

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft, J. (2017). Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 349–354.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Goga, O., Venkatadri, G., and Gummadi, K. P. (2015). The doppelgänger bot attack: Exploring identity impersonation in online social networks. In *Proceedings of the 2015 internet measurement conference*, pages 141–153.

Gong, N. Z. and Liu, B. (2016). You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 979–995.

Gong, N. Z. and Liu, B. (2018). Attribute inference attacks in online social networks. *ACM Transactions on Privacy and Security (TOPS)*, 21(1):1–30.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Guarino, S., Trino, N., Celestini, A., Chessa, A., and Riotta, G. (2020). Characterizing networks of propaganda on twitter: a case study. *Applied Network Science*, 5(1):1–22.

Guo, Q., Xie, H., Li, Y., Ma, W., and Zhang, C. (2021). Social bots detection via fusing bert and graph convolutional networks. *Symmetry*, 14(1):30.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., and Zadeh, R. (2013). Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hand, D. J. and Yu, K. (2001). Idiot's bayes—not so stupid after all? *International statistical review*, 69(3):385–398.

Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. (2012). Preventing private information inference attacks on social networks. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1849–1862.

Hu, Z., Dong, Y., Wang, K., and Sun, Y. (2020). Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, pages 2704–2710.

Jain, N., Agarwal, P., and Pruthi, J. (2015). Hashjacker-detection and analysis of hashtag hijacking on twitter. *International journal of computer applications*, 114(19).

Jia, J., Wang, B., Zhang, L., and Gong, N. Z. (2017). Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1561–1569.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. (2022). Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Jun, Y., Meng, R., and Johar, G. V. (2017). Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences*, 114(23):5976–5981.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., and Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20(3):531–558.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM computing surveys (CSUR)*, 31(4es):5–es.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., and Baddour, K. (2020). Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3).

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections*, 24(3):43–52.

Kumar, S., Zhang, X., and Leskovec, J. (2019). Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1269–1278.

Le, H., Boynton, G., Shafiq, Z., and Srinivasan, P. (2019). A postmortem of suspended twitter accounts in the 2016 us presidential election. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265. IEEE.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520.

Lee, K., Caverlee, J., and Webb, S. (2010). Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442.

Lee, K., Eoff, B., and Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 185–192.

Lingam, G., Rout, R. R., and Somayajulu, D. V. (2019). Adaptive deep q-learning model for detecting social bots and influential users in online social networks. *Applied Intelligence*, 49(11):3947–3964.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martinelli, F., Mercaldo, F., and Santone, A. (2019). Social network polluting contents detection through deep learning techniques. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., and Tesconi, M. (2019). Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Mei, B., Xiao, Y., Li, H., Cheng, X., and Sun, Y. (2017). Inference attacks based on neural networks in social networks. In *Proceedings of the fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies*, page 10. ACM.

Milon-Flores, D. F. and Cordeiro, R. L. (2022). How to take advantage of behavioral features for the early detection of grooming in online conversations. *Knowledge-Based Systems*, 240:108017.

Mora-Cantallops, M., Sánchez-Alonso, S., and Visvizi, A. (2019). The influence of external political events on social networks: The case of the brexit twitter network. *Journal of Ambient Intelligence and Humanized Computing*, 12:4363–4375.

Mulders, D., De Bodt, C., Bjelland, J., Pentland, A., Verleysen, M., and de Montjoye, Y.-A. (2019). Inference of node attributes from social network assortativity. *Neural Computing and Applications*, pages 1–21.

Narayan, K., Agarwal, H., Mittal, S., Thakral, K., Kundu, S., Vatsa, M., and Singh, R. (2022). Desi: Deepfake source identifier for social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2858–2867.

Newman, M. (2010). *Networks: an introduction*. Oxford university press.

Newman, M. E. (2016). *Mathematics of networks*. Springer.

Noulas, A., Mascolo, C., and Frias-Martinez, E. (2013). Exploiting foursquare and cellular data to infer user activity in urban environments. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 167–176. IEEE.

Opitz, J. and Burst, S. (2019). Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251.

OSINT (2019). Twint. `https://github.com/twintproject/twint`.

OSoMe (2022). Bot repository.

Pacheco, D., Flammini, A., and Menczer, F. (2020a). Unveiling coordinated groups behind white helmets disinformation. In *Companion Proceedings of the Web Conference 2020*, pages 611–616.

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Tran Truong, B., Flammini, A., and Menczer, F. (2020b). Uncovering coordinated networks on social media. *arXiv*.

Pastor-Galindo, J., Zago, M., Nespoli, P., Bernal, S. L., Celdrán, A. H., Pérez, M. G., Ruipérez-Valiente, J. A., Pérez, G. M., and Mármol, F. G. (2020). Spotting political social bots in twitter: A use case of the 2019 spanish general election. *IEEE Transactions on Network and Service Management*, 17(4):2156–2170.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Popper, K. R. (1971). *The open society and its enemies: The spell of Plato*, volume 1. Princeton University Press.

Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Potha, N. and Stamatatos, E. (2018). Intrinsic author verification using topic modeling. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–7.

Preoţiuc-Pietro, D. and Ungar, L. (2018). User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th international conference on computational linguistics*, pages 1534–1545.

Rathore, S., Sharma, P. K., Loia, V., Jeong, Y.-S., and Park, J. H. (2017). Social network security: Issues, challenges, threats, and solutions. *Information sciences*, 421:43–69.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252.

Rauchfleisch, A. and Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PloS one*, 15(10):e0241045.

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A. R., and Stamatatos, E. (2016). Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33.

Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011). Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer.

Rosenberg, H., Syed, S., and Rezaie, S. (2020). The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic. *Canadian journal of emergency medicine*, 22(4):418–421.

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., and Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sánchez-Corcuera, R., Bilbao-Jayo, A., Zulaika, U., and Almeida, A. (2021a). Analysing centralities for organisational role inference in online social networks. *Engineering Applications of Artificial Intelligence*, 99:104129.

Sánchez-Corcuera, R., Zubiaga, A., and Almeida, A. (2021b). Analyzing the existence of organization specific languages on twitter. *IEEE Access*, 9:111463–111471.

Sánchez-Corcuera, R., Zubiaga, A., and Almeida, A. (2022). Achieving participatory smart cities by making social networks safer. In *Proceedings of the 7th International Conference on Smart and Sustainable Technologies (SpliTech 2022)*, pages 1–6.

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2020). Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2725–2732.

Severyn, A. and Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962.

Shalev-Shwartz, S. and Ben-David, S. (2014). Decision trees. In *Understanding Machine Learning*, pages 250–256. Cambridge university press.

Sharma, N. (2022). What will elon musk tweet next: Generating tweets with deep learning – confusedcoders.

Singh, A., Blanco, E., and Jin, W. (2019). Incorporating emoji descriptions improves tweet classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2096–2101.

Sippey, M. (2012). Changes coming in version 1.1 of the twitter api.

Skorniakov, K., Turdakov, D., and Zhabotinsky, A. (2018). Make social networks clean again: Graph embedding and stacking classifiers for bot detection. In *CIKM Workshops*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stamatatos, E. (2017). Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149.

Stieglitz, S., Brachten, F., Ross, B., and Jung, A.-K. (2017). Do social bots dream of electric sheep? a categorisation of social media bot accounts. *arXiv preprint arXiv:1710.04044*.

Suarez-Lledo, V., Alvarez-Galvez, J., et al. (2021). Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187.

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., and Menczer, F. (2016). The darpa twitter bot challenge. *Computer*, 49(6):38–46.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Sánchez-Corcuera, R. (2022a). rubensancor/foreseeing_actions_data: v1.0.0. https://doi.org/10.5281/zenodo.6862544.

Sánchez-Corcuera, R. (2022b). rubensancor/organisation_language_twitter: v1.0.0. https://doi.org/10.5281/zenodo.6862495.

Sánchez-Corcuera, R. (2022c). rubensancor/rolemining_twitter: v1.0.0. https://doi.org/10.5281/zenodo.6865869.

Twitter (2020). Insights into attempts to manipulate twitter by state-backed entities.

Twitter (2022).

Valverde-Rebaza, J. C., Roche, M., Poncelet, P., and de Andrade Lopes, A. (2018). The role of location and social strength for friendship prediction in location-based social networks. *Information Processing & Management*, 54(4):475–489.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*, pages 280–289.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Wang, A. H. (2010). Don't follow me: Spam detection in twitter. In *2010 international conference on security and cryptography (SECRYPT)*, pages 1–10. IEEE.

Welling, M. and Kipf, T. N. (2016). Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Wu, B., Liu, L., Yang, Y., Zheng, K., and Wang, X. (2020). Using improved conditional generative adversarial networks to detect social bots on twitter. *IEEE Access*, 8:36664–36680.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, C., Harkreader, R., and Gu, G. (2013). Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293.

Yang, K.-C., Ferrara, E., and Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *arXiv preprint arXiv:2201.01608*.

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., and Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61.

Yang, K.-C., Varol, O., Hui, P.-M., and Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103.

Yardi, S., Romero, D., Schoenebeck, G., et al. (2010). Detecting spam in a twitter network. *First monday*.

Yuan, X., Schuchard, R. J., and Crooks, A. T. (2019). Examining emergent communities and social bots within the polarized online vaccination debate in twitter. *Social media+ society*, 5(3):2056305119865465.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*, pages 218–226.

Zarei, K., Farahbakhsh, R., and Crespi, N. (2019a). Deep dive on politician impersonating accounts in social media. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

Zarei, K., Farahbakhsh, R., and Crespi, N. (2019b). Typification of impersonated accounts on instagram. In *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*, pages 1–6. IEEE.

Zarrinkalam, F., Kahani, M., and Bagheri, E. (2018). Mining user interests over active topics on social networks. *Information Processing & Management*, 54(2):339–357.

Zhan, J., Gurung, S., and Parsa, S. P. K. (2017). Identification of top-k nodes in large networks using katz centrality. *Journal of Big Data*, 4(1):1–19.

Zhang, J., Hu, X., Zhang, Y., and Liu, H. (2016a). Your age is no secret: Inferring microbloggers' ages via content and interaction analysis. In *Tenth International AAAI Conference on Web and Social Media*, pages 476–485. IEEE.

Zhang, J., Zhang, R., Zhang, Y., and Yan, G. (2016b). The rise of social botnets: Attacks and countermeasures. *IEEE Transactions on Dependable and Secure Computing*, 15(6):1068–1082.

Zhang, X. and Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

Zhang, Y. and Wallace, B. C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263.

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., and Rong, C. (2015). Detecting spammers on social networks. *Neurocomputing*, 159:27–34.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

APPENDIX

A

Hyperparameter optimisation for language classification model

We performed a hyperparameter optimisation to get the best results in the model proposed in Section 3.2 for language classification. For this purpose, we use the Sweep tool offered by the Weight and Biases (WandB) platform. WandB is a platform dedicated to managing AI projects that offer services from evaluating the models to creating pipelines to deploy models in production.

The Sweep tool offered by this platform allows us to run evaluations by continuously changing the parameters of the models to find the best ones. In this case, we run a Bayesian search to model the relationships between the parameters and optimise a metric, the F-score. Each evaluation was conducted using a 5-fold cross-validation method to ensure that the hyperparameters were not optimised for a specific dataset.

Figure shows the F-score obtained for each evaluation as well as the studied parameters:

Parameter	Value
Batch	1007
Dense size 1	849
Dense size 2	242
Dropout	0.537390285717924
Learning rate	0.001816368228041
Kernel start	8
Kernel step	2

Table A.1: Best parameters found in the Sweep for the model proposed in Section 3.2.

- **Batch:** Batch size is the number of samples that are processed before the model is updated.
- **Dense size:** Output size of the dense layer. We set different sizes for each dense layer.
- **Dropout:** Dropout randomly sets input units to 0 during the training step with a frequency to which this parameter is set.
- **Learning rate:** This parameter sets the rate at which the weights are updated during the backpropagation phase used by the stochastic gradient optimisation algorithm.
- **Kernel start:** This parameter sets the dimension at which the convolutional layers employed in the model will start.
- **Kernel steps:** This parameter sets how many convolutional layers the model will implement. The convolutional layers will start at the dimension set by the Kernel start parameter and will increase one at a time.

Finally, Table A.1 shows the best parameters found for the proposed model.

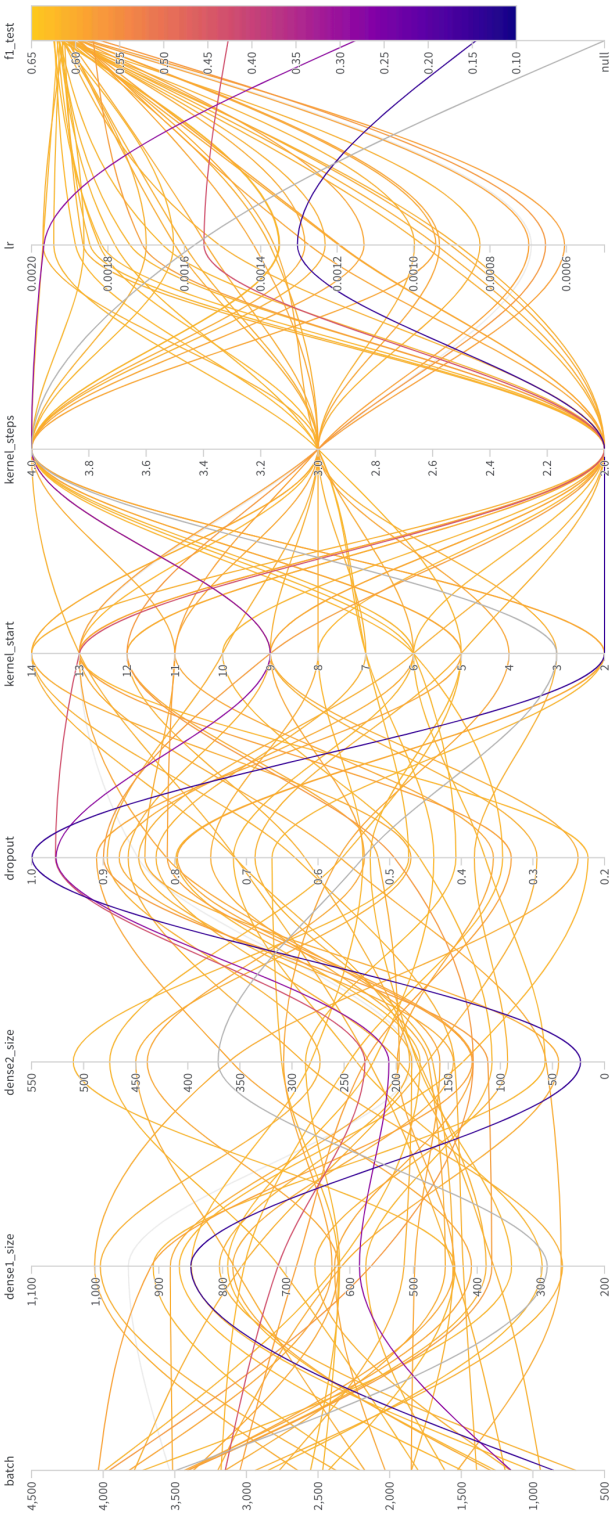

Figure A.1: Results of the evaluations conducted in the Sweep.

Declaration

I, Rubén Sánchez Corcuera, here with declare that this dissertation is my own original work, carried out as a doctoral student at the University of Deusto. All assistance received and notions from other sources have been identified as such, acknowledging their correspondent contributions and citing them properly.

This work contains no material which has been presented in identical or similar form to any examination board, except where due acknowledgement is made in the dissertation.

This dissertation was finished writing on September 9^{th}, 2022