future internet

MDPI

Article

Cyberbullying Detection, Prevention, and Analysis on Social Media via Trustable LSTM-Autoencoder Networks over Synthetic Data: The TLA-NET Approach †

Alfredo Cuzzocrea [1,2,*,‡], Mst Shapna Akter [3], Hossain Shahriar [4] and Pablo Garcia Bringas [5]

[1] iDEA Lab, University of Calabria, 87036 Rende, Italy
[2] Department of Computer Science, University of Paris City, 75013 Paris, France
[3] Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA; akter@oakland.edu
[4] Center for Cybersecurity, University of West Florida, Pensacola, FL 32514, USA; hshahriar@uwf.edu
[5] Faculty of Engineering, University of Deusto, 48007 Bilbao, Bizkaia, Spain; pablo.garcia.bringas@deusto.es
* Correspondence: alfredo.cuzzocrea@unical.it

† This paper is the extension of our paper: Akter, Mst. S., Shahriar, H., Cuzzocrea, A., Wu, F., Rodriguez-Cardenas, J. "A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data". In Proceedings of 2023 IEEE International Conference on Big Data, Sorrento, Italy, 15–18 December 2023. https://doi.org/10.1109/BigData59044.2023.10386719.

‡ This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France.

Abstract: The plague of cyberbullying on social media exerts a dangerous influence on human lives. Due to the fact that online social networks continue to daily expand, the proliferation of hate speech is also growing. Consequentially, distressing content is often implicated in the onset of depression and suicide-related behaviors. In this paper, we propose an innovative framework, named as the trustable `LSTM-autoencoder` network (`TLA NET`), which is designed for the detection of cyberbullying on social media by employing synthetic data. We introduce a state-of-the-art method for the automatic production of translated data, which are aimed at tackling data availability issues. Several languages, including Hindi and Bangla, continue to face research limitations due to the absence of adequate datasets. Experimental identification of aggressive comments is carried out via datasets in Hindi, Bangla, and English. By employing `TLA NET` and traditional models, such as long short-term memory (`LSTM`), bidirectional long short-term memory (`BiLSTM`), the `LSTM-autoencoder`, `Word2vec`, bidirectional encoder representations from transformers (`BERT`), and the Generative Pre-trained Transformer 2 (`GPT-2`), we perform the experimental identification of aggressive comments in datasets in Hindi, Bangla, and English. In addition to this, we employ evaluation metrics that include the F1-score, accuracy, precision, and recall, to assess the performance of the models. Our model demonstrates outstanding performance across all the datasets by achieving a remarkable 99% accuracy and positioning itself as a frontrunner when compared to previous works that make use of the dataset featured in this research.

Keywords: cyber-bullying; deep learning; neural networks; natural language processing

Academic Editors: Steven Furnell and Wei Yu

Received: 16 June 2024
Revised: 11 December 2024
Accepted: 25 January 2025
Published: 12 February 2025

Citation: Cuzzocrea, A.; Akter, M.S.; Shahriar, H.; Garcia Bringas, P. Cyberbullying Detection, Prevention, and Analysis on Social Media via Trustable LSTM-Autoencoder Networks over Synthetic Data: The TLA-NET Approach. *Future Internet* **2025**, *17*, 84. https://doi.org/10.3390/fi17020084

1. Introduction

The World Wide Web (`WWW`) plays a pivotal role in the substantial growth of social media platforms [1,2], like Facebook, Twitter, Instagram, and YouTube. These platforms enable users to share a vast array of information, leading to increased user interactions. However,

this also leads to an unregulated surge in online hate speech [3], a phenomenon that, in extreme cases, can drive individuals to take their own lives [4]. One of the most concerning phenomena is cyberbullying, which entails the dissemination of offensive content or other forms of violence through digital media [5–7] with the intent to harm an individual or a group of individuals. In particular, teenagers can assume the roles of victims, perpetrators, or simply bystanders [8] when it comes to cyberbullying. One study revealed that 36.5% of students have experienced cyberbullying at least once in their lives. Among all types of online comments, rude or cruel ones are the most prevalent. In a study conducted in the United States, which involved 1501 adolescents aged 10 to 17, it was discovered that 12% of the sample admitted to engaging in mistreatment, 4% declared themselves as victims, and 3% acknowledged being both aggressors and victims of cyberbullying [9].

The Cybercrime Division (`CID`) of Sri Lanka recently conducted a survey and detected that over 1000 cases of cyberbullying have been recorded. Surprisingly, more than 90% of university students reported being victims of cyberbullying, while nearly all survey participants stated that they knew someone who had experienced online bullying. It is alarming to note that 80% of the reported cases of cyberbullying in Sri Lanka occurred on Facebook, with 65% of university students admitting to sharing embarrassing videos or photos. Furthermore, the survey reveals that 15% of users share personal information online and 9% spread inaccurate information and falsehoods about others, while only 2% post offensive material [10].

One of the main problems of cyberbullying is its rapid spread due to the large online audience. Due to the fact that phenomenon has become a daily occurrence, victims often suffer severe consequences [11]. Based on the MetroWest Adolescent Health Survey, Schneider et al. [12] present the relationship between victimization and five categories of psychological distress. Information was collected from over 20,000 students. Among cyberbullying victims, self-harm (24%) and depressive symptoms (34%) showed the highest rates of psychological distress. These findings highlight a concerning connection between cyberbullying and psychological distress, underscoring the need to address this issue urgently.

An exponential increase in social media users and the subsequent rise of the cyberbullying phenomenon require a thorough investigation. However, previous studies (e.g., [13–18]) have addressed the issue using inefficient algorithms and with a limited amount of data for training AI algorithms. In actuality, studies conducted so far for the detection of the cyberbullying phenomena of social media present some significant limitations. For instance, Perera et al. [10] used a limited number of data instances (1000 labeled textual data) for classification by employing classifiers, like a Support Vector Machine (`SVM`) [19], but achieved an accuracy of 74%, which is considered inadequate. Alotaibi et al. [20] proposed a multicamera deep learning framework based on a dataset of 55,788 tweets and developed a multicamera deep learning approach, but the 88% accuracy rate could still be improved upon. Ahmed et al. [21] used a Deep Neural Network with 44,001 comments from Facebook and attempted to develop a Hybrid Neural Network, but they achieved an accuracy of only 85%, which may not be deemed acceptable.

Therefore, a robust and efficient model is required. In this research, we have developed a neural network model that provides highly accurate results, called the trustable `LSTM-autoencoder` network (`TLA-NET`). Furthermore, our model is capable of ensuring very high accuracy, overcoming the tendency of other methods to produce inaccurate results in the final layer. Specifically, we address issues related to the lack of available data, such as the absence of datasets for recognizing aggressive comments in Bangla and Hindi. For devising our technique, we utilize the TRAC-2 dataset (Workshop on Trolling, Aggression, and Cyberbullying) [22], which encompasses comments in English, Bangla, and Hindi. Furthermore, we create a dataset that is entirely translated automatically into English to tackle

data accessibility challenges. Our approach to linguistic translation is straightforward; we employ Google Translate, which is easily accessible and free. While Google Translate translates data between languages, it introduces a certain level of noise that may not suit training deep learning models. We refer to this synthetic data as noisy since it undergoes machine translation and contains various sources of noise. It is important to notice here that, while the adoptation of Google Translate introduces, of course, a certain degree of noise, thus reducing the accuracy of the overall dataset, in our experimental work we applied a procedure to ensure the highest possible accuracy of the dataset to be used in our experiments. This approach, while valuable for our actual research, is poorly automatically generalizable in other languages. However, for each specific case, pragmatic solutions, like ours, can be investigated.

In this study, we aim to determine whether the proposed deep learning model can identify patterns in noisy data. Looking at the details, our `TLA-NET` model employs language transformers, such as bidirectional encoder representations from transformers (`BERT`) [23] and the Generative Pre-trained Transformer 2 (`GPT-2`) [24], along with simple neural networks, like long short-term memory (`LSTM`) [25], bidirectional long short-term memory (`BiLSTM`) [26], the `LSTM-autoencoder` [27–29], and `Word2Vec` [30].

In our experimental study, we considered the reference Bengali Cyberbullying Text Dataset [31], which provides a reliable testbed for performance evaluation and assessment. Also, our comparison is based on standard evaluation metrics, including accuracy, precision, recall, and F1-score [32], and advanced evaluation metrics [32], including mean absolute error (`MAE`), mean squared error (`MSE`), and root mean square error (`RMSE`), plus other components, like confusion matrices [32] and receiver operating characteristic (`ROC`) curves [32]. We also consider, as competitive algorithms, some well-positioned-in-the-community machine learning (`ML`) algorithms, like the decision tree (`DT`) [33], random forest (`RF`) [34], logistic regression (`LR`) [35], and multilayer perceptron (`MLP`) models [36].

The proposed `TLA-NET` model delivers state-of-the-art results with an accuracy of about 99% in the raw Bangla dataset. It also achieves top-notch performance despite the presence of partially noisy datasets. Therefore, our method can be valuable for languages with limited data availability. Following our approach, efforts to reduce suicide-related incidents can be somewhat aided through the detection of aggressive comments and the implementation of prediction-based measures (e.g., [37]).

On the other hand, collecting social media data raises the very relevant problem of ensuring the privacy of users (e.g., [38,39]). While this problem is of evident significance, it is not addressed in our research, and is left for future work to consider.

All considered, the primary contributions of this article can be summarized in three aspects, as follows:

1. `TLA-NET`: a trustable `LSTM`-autoencoder network for cyberbullying detection and prevention on social media using synthetic data. The model follows the basic `LSTM-autoencoder` framework and is capable of delivering highly accurate results. We achieve state-of-the-art results using our proposed model.
2. We outline a process to generate synthetic data, thus achieving the so-called noisy synthetic dataset.
3. We conduct extensive experiments by comparing `TLA-NET` against state-of-the-art `ML` algorithms on top of a reference dataset on Bangli cyberbullying text and adopting several authoritative evaluation metrics. This step proves essential for comparing various neural network models under varying levels of noise intensity in the dataset.

This paper is an extension of our previous paper [40], where we provided the initial ideas about the framework and a preliminary classification algorithm as a first prototype.

The remaining part of this paper is organized as follows. Section 2 focuses on reviewing several related works in the context of cyberbullying detection, prevention, and analysis on social media. In Section 3, we examine and describe the methodological approach that we adopted to investigate the phenomenon of cyberbullying within social networks, by presenting our proposed TLA-NET model and the datasets used in our experimental campaign. Section 4 presents the main algorithm implemented by our technique. After that, Section 5 provides a detailed overview of the results obtained through the rigorous application of our proposed TLA-NET model. Finally, Section 6 provides the conclusions of our research and sketches possible future directions.

2. Related Work

In this section, we provide a comprehensive overview of state-of-the-art proposals and relevant related work in the context of cyberbullying detection, prevention, and analysis on social media.

Ref. [41] focuses on addressing the annoying problem of detecting aggressive activity and cyberbullying behavior in textual content on social media platforms, which is critical for reducing its harmful impacts. Existing approaches leveraging traditional natural language processing (NLP) and ML techniques often face limitations in accuracy and contextual understanding. To address these drawbacks, this paper proposes an automated deep learning model using a binary chimp optimization (BCO)-based feature selection (BCO-FSS) technique and a stacked bidirectional gated recurrent unit (SBiGRU) with attention to spatial and sequential semantic learning. The study employs datasets from various platforms, like Formspring, Instagram, and MySpace, and integrates linguistic complexity metrics, such as feature density (FD), and finally utilizes BERT as a baseline classifier. The experimental evaluation of the proposed approach achieves high performance in terms of several well-known metrics, such as accuracy, precision, recall, and F1-score. Finally, this approach provides insights into the decision-making process of the model, due to the incorporation of attention mechanisms that visualize how sentiment weights are assigned at different layers.

Ref. [42] introduces an innovative and reliable framework for detecting cyberbullying built on top of a multi-modal cyberbullying detection (MMCD) approach that integrates data from diverse social networks. The proposed model effectively addresses the limitations of traditional text-based detection methods in handling the multifaceted nature of cyberbullying, by integrating hierarchical attention networks (HANs) with multi-modal data, including visual, textual, and temporal features, in order to provide a holistic cyberbullying detection platform. Specifically, it incorporates a self-attention-based BiLSTM model and HAN model focusing on word and comment levels, and other embedding techniques. The integration of these components enables efficient information merging and contributes to the effective resolution of the complex issue of cyberbullying. Moreover, the experimental analysis conducted over three social media datasets highlights that the MMCD model achieves greater performance when compared with several state-of-the-art models in terms of accuracy and F1-score.

In Ref. [43], the authors propose RoBERTaNET, an enhanced RoBERTa transformer model tailored for cyberbullying detection with global vectors for word representation (GloVe) word embedding features. Through this innovative approach, this research deals with the challenge of automatically identifying cyberbullying in Tweets from a publicly available cyberbullying dataset. The experimental analysis involves comparing the proposed approach with several state-of-the-art machine learning (e.g., random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN)), deep learning (e.g., bidirectional long short-term memory (BiLSTM), convolutional long short-term memory (ConvLSTM)), and

transformer-based models (e.g., bidirectional encoder representations from transformers (BERT)). The results demonstrate the effectiveness of RoBERTaNET by achieving a notable 95% accuracy, along with 95%, 97%, and 96% for precision, recall, and F1-score, respectively. In summary, this approach of combining robust transformer-based architectures with word embedding techniques paves the way for more effective and scalable solutions.

Ref. [38] proposes FedBully, a novel federated learning (FL) approach for cyberbullying detection to preserve client privacy in learning systems. The proposed approach leverages sentence encoders for feature extraction and secure aggregation protocols in order to ensure data privacy while enabling effective collaborative learning. Therefore, the contributions of this research are threefold: (i) proposing FedBully for fast and privacy-preserving cyberbullying detection by adapting secure aggregation; (ii) demonstrating that secure aggregation protocols do not compromise model performance through extensive evaluations; (iii) conducting detailed analysis on the sampling rate and dataset imbalance to provide insights into the model sensitivity. Furthermore, an extensive experimental evaluation is conducted to optimize FedBully hyperparameters, in addition, it should be noted that the model achieves promising classification performance, with an over 93% of AUC metric over public large-scale cyberbullying datasets.

In Ref. [44], the authors present Bengalibullying, a robust hybrid machine learning model for cyberbullying detection in the Bengali language on social media. The proposed approach incorporates effective text preprocessing, TfidfVectorizer (TF-IDF)-based feature extraction, and instance hardness threshold (IHT) resampling to address dataset imbalance and improve model performance. Moreover, this research performs a comparative analysis with several ML models by using state-of-the-art metrics (e.g., accuracy, recall, F1-score, etc.). Using a publicly available dataset of 44,001 Bengali comments, the model achieves state-of-the-art results, with 98.57% accuracy in binary classification and 98.82% in multi-label classification, which outperforms all prior efforts in this domain. This research demonstrates the model's potential for real-life application in cyberbullying detection systems, and offers a powerful tool for fighting cyberbullying across multiple language contexts.

3. Methodology

In this section, we carefully examine the methodological approach adopted to investigate the phenomenon of cyberbullying within social networks. The methodology not only provides the framework through which data are collected and analyzed but also emphasizes the accuracy and validity of the results obtained. This section outlines in detail each step taken to ensure a rigorous and scientifically sound approach to our work.

3.1. Dataset Specification

In this research, the TRAC-2 dataset [22] is selected based on specific criteria to ensure its relevance and suitability for evaluating our proposed model performance in multilingual contexts. This dataset includes labeled comments in English, Hindi, and Bangla, which provides a valuable resource for analyzing cyberbullying detection across languages with limited data availability. The inclusion of aggressive, misogynistic (i.e., gender-discriminatory), and non-aggressive labels further aligns with our model focus on identifying harmful online content in diverse linguistic settings.

The dataset used to describe our complex approach comes from TRAC-2 [22]. It comprises 25,000 comments from three social media platforms, namely Facebook, YouTube, and Twitter, in three languages, namely English, Bengali, and Hindi [45]. More specically, according to the guidelines provided by [22], since comments are, of course, of different sizes, in order to have equal-sized comments, string padding has been used, so that

the final comments' lengths were 30, 38, and 33 words for English, Hindi, and Bangla, respectively. On the other hand, the longest comments' lengths were 58, 72, and 62 words for English, Hindi, and Bangla, respectively. The shared task includes two groups: Sub-task A (aggressive comments) and Sub-task B (misogynistic comments). Sub-task A consists of three classes: (*i*) no comments are annotated as covertly aggressive (`CAG`), (*ii*) the direct aggressive comments are annotated as overtly aggressive (`OAG`), and (*iii*) non-aggressive comments are annotated as non-aggressive (`NAG`). Similarly, Sub-task B contains two classes: `GEN` and `NGENA`. A comment that indicates a man, woman, or transgender individual is annotated as `GEN`, and a comment that does not indicate gender is annotated as `NGEN`. In particular, for the sake of simplicity, we used all the comments for our analysis, considering both the emotionally problematic parts and the safe parts of comments. Clearly, a finer analysis could be devised, such as extracting only the emotionally problematic parts from comments. This research aspect, while very interesting, is outside the scope of our paper, and is left for future work. All three datasets contain both training and test sets. In our project, we conducted experiments on Sub-task A, as the features of Sub-task A align perfectly with our goal of prediction, which is cyberbullying detection. The statistics of the dataset provided by the organizations are displayed in Table 1.

To provide more details, some examples of the text data stored in the target dataset are shown in Figure 1.

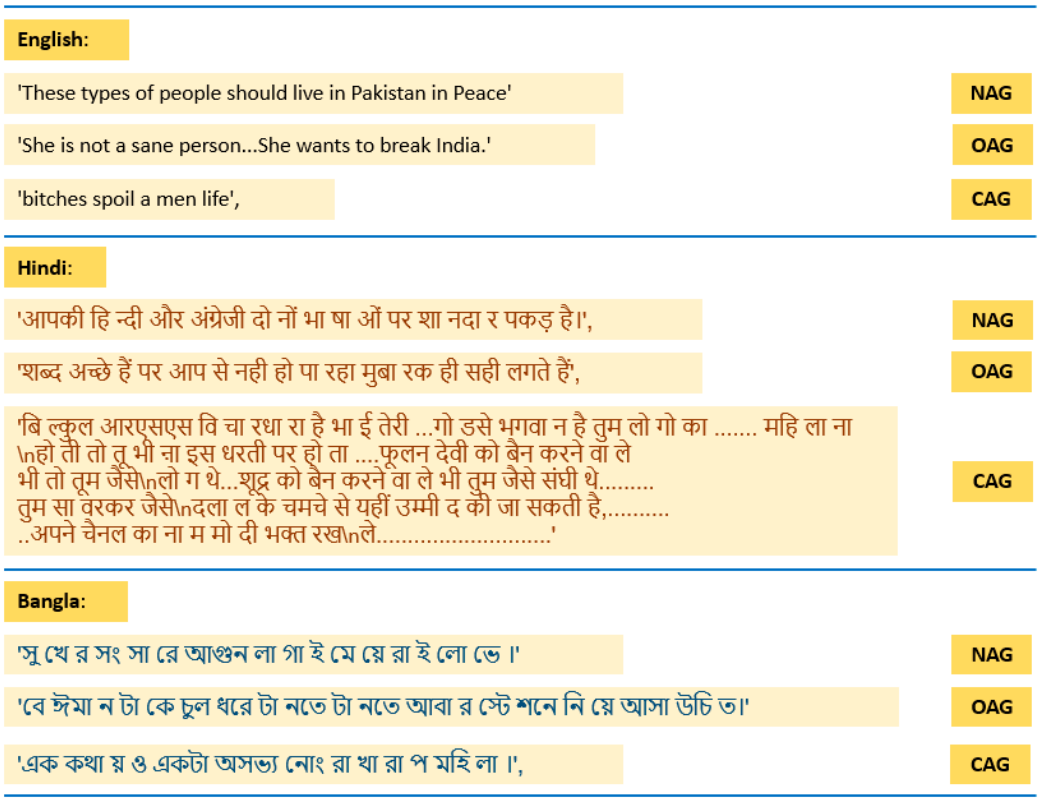


Figure 1. Example of text data in the target dataset.

Table 1. Label distribution in the target dataset for Sub-task A classification.

Author	Data Source	Instance	Method	Limitation
Set	**NAG**	**OAG**	**CAG**	**Total**
English Training	3375	435	435	4263
English Testing	836	117	113	1066

Table 1. *Cont.*

Author	Data Source	Instance	Method	Limitation
Set	**NAG**	**OAG**	**CAG**	**Total**
Hindi Training	2245	823	910	3984
Hindi Testing	578	211	208	997
Bangla Training	2078	898	850	3826
Bangla Testing	522	218	217	957

3.2. Dataset Preprocessing and Augmentation

To create semi-noisy data, we add noise to the raw data to the dataset in order to address imbalance issues. The initial data we use are highly imbalanced for Sub-task A. The imbalanced data performs poorly when predicting aggressive comments. The NAG category comprises 50% of the total text data, while the OAG and CAG categories make up the other 50% of the total text data. To resolve the imbalanced data issue, we augment the text data in a manner that ensures that all classes have nearly the same amount of text data. We employ two methods for the augmentation process: noise addition and data translation.

Noise Addition. We add noise by replacing words with synonyms or antonyms, inserting random stop words and rearranging certain words in the raw text data. This process expands the corpus size while maintaining the original context of the dataset.

Data Translation. Data undergo translation from one language to another, such as from English to Bangla, utilizing Google Translate, as already mentioned. All languages encompass identical sub-tasks and classes. Consequently, we perform translations across all feasible combinations until the dataset attains equilibrium. We have translated texts for the `NAG`, `OAG`, and `CAG` classes within Sub-task A.

We add texts from the noise augmentation process and texts from the translation augmentation process into the raw data so that the new corpus contains an almost equal number of text data for Sub-task A. The dataset's statistics, after incorporating the augmented data with the raw data, are presented in Table 2.

Table 2. Label Distribution in the Augmented Raw Dataset.

Set	NAG	OAG	CAG	Total
English Training	3375	2251	2546	8172
English Testing	836	117	113	1066
Hindi Training	2245	3497	1810	7552
Hindi Testing	578	211	208	997
Bangla Training	2078	1959	1966	6003
Bangla testing	522	218	217	957

3.3. Fully Machine-Translated Data

Through the translation augmentation process, we generate a comprehensive, fully noisy machine-translated English dataset. We translate the Sub-task A texts from Bangla and Hindi into English to create the new dataset. The statistics of the entire translated data are presented in Table 3.

Table 3. Label distribution in the fully translated English dataset.

Set	NAG	OAG	CAG	Total
English Training	4373	3096	2588	10,057

Some examples of the fully-noisy machine-translated English data are shown in Table 4.

Table 4. Example of fully-noisy machine-translated English dataset with categories.

Text	Set
It doesn't matter if Moumita Sarkar or Hindu Der Modda is extramarrief Kora. Well said, brother.	NAG
What are you looking at? 200 rupees has been said to be 1000 rupees, there is a mobile phone, সব n Green has come face to face. You read everything from your reading, কি ছু n don't want to do anything? Here is what we have said before Duran Diacho Sothi Ranu Di is doing very wrong.	NAG
Kunfu coming from China and kutta coming from Pakistan.	OAG
Disliked n unsubscribe. Ohh wait, I wasn't subscribed.	OAG
I hat ranu Mondal.	CAG
One word for all haters to die somewhere else...	CAG

The raw text, semi-noisy, and fully-noisy datasets are converted into a machine-understandable numerical representation. Computers exclusively comprehend numbers, making it essential to convert text into numerical formats before inputting it into `ML` models. We employ a TensorFlow.Keras tokenizer for the `TLA-NET` model.

3.4. Proposed Model

We create a model using an auto-encoder and LSTM for each comment. Figure 2 represents the suggested modified LSTM-autoencoder. Here, an auto-encoder is constructed using the LSTM network; hence, we label it as the LSTM-autoencoder network. Comments serve as the input for this LSTM-autoencoder; then, they are encoded, decoded, and reconstructed to their original forms. The network aims to minimize the distance between the reconstructed representation and the original input as it learns. Our initial training of a standard auto-encoder, which is an auto-encoder with only basic neural layers, resulted in very poor performance. The default auto-encoder treats each word as a separate input and does not consider the order or sequence in which they appear in a comment text. Consequently, there was significant loss during the training phase because comment inputs are sequential texts and the LSTM network is specifically designed to handle such sequences of input data. Therefore, we chose the LSTM-autoencoder as the foundational concept for our proposed model.

The trustable `LSTM-autoencoder` network depicted in Figure 2 can be broadly divided into five subparts: (i) a stacked `LSTM` encoder, (ii) repeated vectors, (iii) a stacked `LSTM` decoder, (iv) reconstruction loss, and (v) the *WisdomNet*. In the following paragraphs, we describe each of these components.

Stacked LSTM encoder. The term stacked `LSTM` network refers to an `LSTM` network that has multiple stacked `LSTM` layers. The network becomes more complex and deep as the `LSTM` layers are stacked. In the stacked `LSTM` encoder part, a lower `LSTM` layer provides multiple outputs to its upper `LSTM` layer, unlike a single output. In other words, it generates an output for each time step of input rather than a single output for all input steps. The last

layer of the LSTM encoder produces a single output vector that contains details about the entire comment.

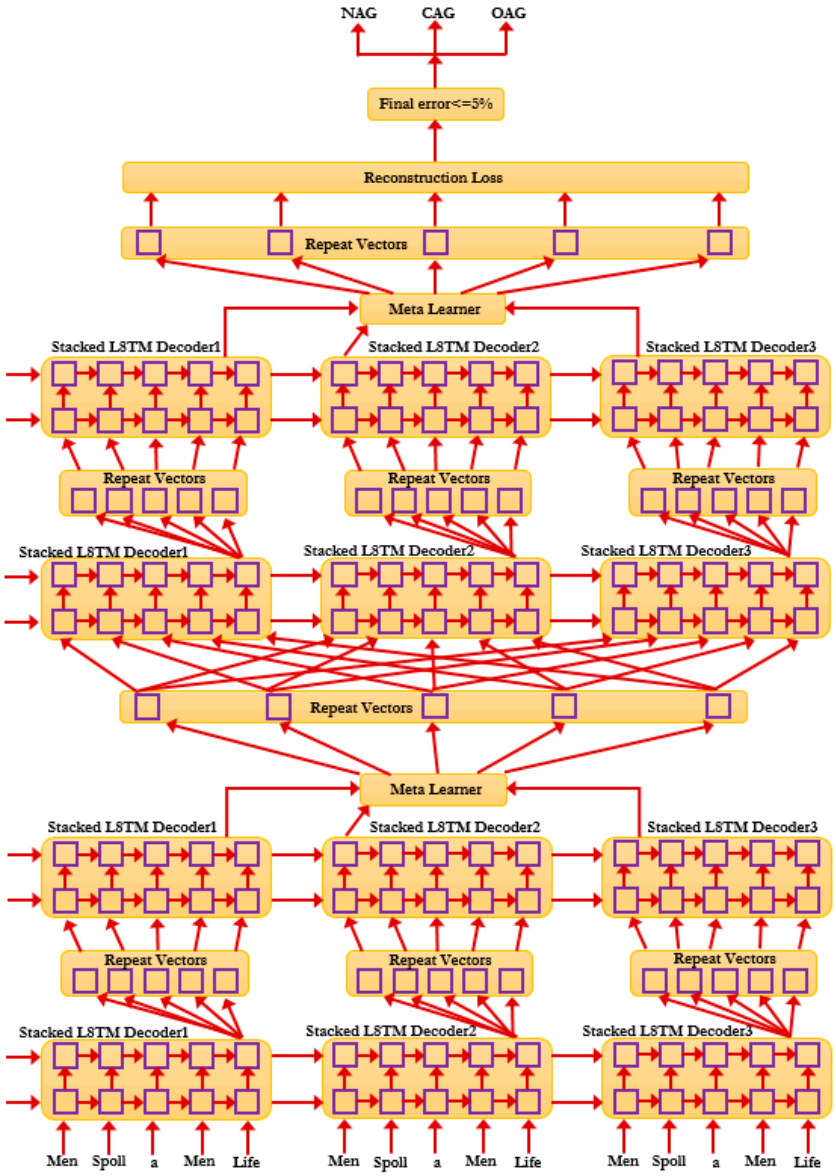


Figure 2. Overview of the modified LSTM-autoencoder network.

Repeat vectors. The next task is to reconstruct the inputs from the encoded input sequence. This means that we have a single output vector from the stacked LSTM encoder to work with in order to iteratively build the input sequence. We use the repeating vectors to repeatedly apply the output vector to the number of time steps in the output sequence to achieve this result.

Stacked LSTM decoder. The stacked LSTM decoder then receives these repeating vectors as an input. This indicates that the input to the stacked LSTM decoder was the same at every time step but had a different hidden state. The original input sequence is reconstructed by stacking several LSTM layers.

Reconstruction loss. The actual output and the expected output for each review $I_i(i_1, i_2, \ldots.., i_t)$ are given as $I_i(i_1, i_2, \ldots.., i_t)$ and $O_i(o_1, o_2, \ldots.., o_t)$, respectively. The reconstruction loss R_{loss} for the review I_i can be evaluated in the following way using mean squared error (MSE):

$$R_{loss} = O_i - I_i, \tag{1}$$

To determine whether a comment is aggressive or not, we calculate the reconstruction loss between the input I_i and the output O_i of the proposed model. We use a simple neural network approach to obtain the threshold. When the reconstruction loss from the auto-encoder is passed through the neural network, each comment is classified as a common aggressive comment (`CAG`), a common non-aggressive comment (`NAG`), or common extremely aggressive comment (`OAG`). For each comment, the neural network returns the group label (`CAG`, `NAG`, or `OAG`). The overall accuracy of the model is then determined by comparing these predicted labels with the actual labels of the comments.

WisdomNet. WisdomNet is a neural network architecture for reliable learning that can determine what cannot be accurately classified. The WisdomNet is constructed in four phases: (i) a standard neural network is trained and misclassified training samples are identified; (ii) a new network is created based on the trained network; (iii) a new neuron is added to the output layer to produce the label "unable to classify"; (iv) the new network is retrained using only the misclassified training samples. Similar to the neuron that returns the class label, the conjugate neuron also produces a rejection or the label unable to classify in this situation.

The proposed model takes each comment in parallel through three stacked `LSTM` encoders (stacked `LSTM encoder 1`, stacked `LSTM encoder 2`, and `stacked LSTM encoder 3`). Each encoder encodes the input and immediately passes it through repeat vectors to reconstruct its shape. Each reshaped encoded output then passes through another stacked `LSTM` encoder. Finally, the outputs from the three stacked `LSTM` encoders are fed into a meta-learner to provide a single encoded output. The meta-learner is a simple `RNN` model that quickly learns from weighted inputs. The single output then passes through a repeat vector to reshape it before feeding into the stacked `LSTM` decoder.

The final outputs of the decoders are fed into a meta-learner, which is a simple neural network designed to provide a single output. This single output is then used for repeat vectors to scale the information, and it is subsequently employed to compute the reconstruction loss. During the final phase of the reconstruction calculation, a neuron called *WisdomNet* is connected to the simple neural network to learn from the misclassified outputs. We iterate the optimization process using gradient descent until we achieve an error greater than 5 and a misclassification rate higher than 5%, ensuring a highly accurate result.

4. Main Algorithm

In this section, we provide the main algorithm implemented by our technique, i.e., Algorithm 1, namely `CyberbullismDetectionAndAnalysis`, which is expressed in Java-inspired pseudo-language.

Algorithm 1. Algorithm CyberbullismDetectionAndAnalysis

```
Algorithm CyberbullismDetectionAndAnalysis
Input Collection of comments from social media.
Output Evaluation of comments: aggressive, non-aggressive, and accuracy,
precision, recall, f1_score, roc_curve, confusion_matrix metrics.
```

```
import org.tensorflow.lite.Interpreter;
import org.tensorflow.lite.gpu.GpuDelegate;
```

Algorithm 1. *Cont.*

```
import org.tensorflow.lite.support.common.FileUtil;
import org.tensorflow.lite.support.model.Model;
import org.tensorflow.lite.support.model.Model.Options;
import java.util.List;
import org.apache.commons.io.FileUtils;
import org.apache.commons.text.similarity.CosineDistance;
import org.deeplearning4j.models.word2vec.WordVectors;
import
org.deeplearning4j.text.tokenization.tokenizer.preprocessor.CommonPreprocessor;
import
org.deeplearning4j.text.tokenization.tokenizer.preprocessor.TokenPreProcess;
import
org.deeplearning4j.text.tokenization.tokenizerfactory.DefaultTokenizerFactory;
import org.deeplearning4j.text.tokenization.tokenizerfactory.TokenizerFactory;
import edu.stanford.nlp.util.Pair;
import org.tensorflow.lite.support.tensorbuffer.TensorBuffer;

public class CyberbullismDetectionAndAnalysis {
private static final String BERT_MODEL_PATH = ‘‘bert_model.tflite’’;
private static final String GPT2_MODEL_PATH = ‘‘gpt2_model.tflite’’;
private static final String LSTM_MODEL_PATH = ‘‘lstm_model.tflite’’;
private static final String BILSTM_MODEL_PATH = ‘‘bilstm_model.tflite’’;
private static final String LSTM_AUTOENCODER_MODEL_PATH =
‘‘lstm_autoencoder_model.tflite’’;
private static final String WORD2VEC_MODEL_PATH = ‘‘word2vec_model.bin’’;
private static final TokenPreProcess TOKEN_PRE_PROCESSOR = new
CommonPreprocessor();
private static final TokenizerFactory TOKENIZER_FACTORY = new
DefaultTokenizerFactory();
private static final CosineDistance COSINE_DISTANCE = new CosineDistance();
private static final float SIMILARITY_THRESHOLD = 0.5f;

// Load the previously trained LSTM model
private static Interpreter modelLSTM = new
Interpreter(FileUtils.loadMappedFile(LSTM_MODEL_PATH));

// Load the previously trained BiLSTM model
private static Interpreter modelBiLSTM = new
Interpreter(FileUtils.loadMappedFile(BILSTM_MODEL_PATH));

// Load the previously trained LSTM-Autoencoder model
private static Interpreter modelLSTMAutoencoder = new
Interpreter(FileUtil.loadMappedFile(LSTM_AUTOENCODER_MODEL_PATH));

// Load the previously trained Word2Vec model
private static WordVectors modelWord2Vec =
WordVectorSerializer.readWord2VecModel(WORD2VEC_MODEL_PATH);

// Class constructor
public static CyberbullismDetectionAndAnalysis() {
```

Algorithm 1. *Cont.*

```
// Implementation of the system setup
}

// Function to label an aggressive comment using BERT
public static int labelCommentBERT(String commentText) {
// Implementation for BERT labelling
return predictedLabel;
}

// Function to label an aggressive comment using GPT-2
public static int labelCommentGPT2(String commentText) {
// Implementation for GPT-2 labeling
return predictedLabel;
}

// Function to label an aggressive comment using LSTM
public static int labelCommentLSTM(String commentText) {
// Implementation for LSTM labelling
return predictedLabel;
}

// Function to label an aggressive comment using BiLSTM
public static int labelCommentBiLSTM(String commentText) {
// Implementation for BiLSTM labelling
return predictedLabel;
}

// Function to label an aggressive comment using LSTM-Autoencoder
public static int labelCommentLSTMAutoencoder(String commentText) {
// Implementation for LSTM-Autoencoder labelling
return predictedLabel;
}

// Function to label an aggressive comment using Word2Vec
public static int labelCommentWord2Vec(String commentText) {
// Implementation for Word2Vec labelling
return predictedLabel;
}

// Function to evaluate a model returning accuracy, precision, recall, f1_score,
roc_curve, confusion_matrix
public static List<Float> evaluateModel(List<Integer> predictedLabel,
List<Integer> realLabel) {
// Implementation for model evaluation
return evaluationMetrics;
}
}
```

As shown by Algorithm 1, our technique exploits well-understood deep learning methods to magnify the process of detecting, preventing, and analyzing cyberbullying on social media.

5. Experimental Analysis and Results

This section provides a detailed overview of the results obtained through the rigorous application of our methodology. Through graphs, statistics, and in-depth interpretations, we unveil the nuances and trends that outline the complex landscape of this digital phenomenon. The results presented here not only address the research questions but also open new perspectives and questions that will enrich the academic and practical discourse on the dynamics of cyberbullying within social networks. This study presents a hybrid machine learning approach for online cyberbullying detection. The developed procedure consists of several processes: (i) text preprocessing to prepare the comments, (ii) feature extraction to transform the processed text into numerical information, and (iii) resampling to balance the data.

5.1. Experimental Setup

To create the proposed model, experiments were conducted by considering the Bengali Cyberbullying Text Dataset [31]. The usage of a specific dataset for the experimental phase improves our analysis. As regards the computational infrastructure, we made use of a workstation running the Microsoft Windows 10 Pro operating system with an Intel(R) Core(TM) i7-10700 CPU at 2.90 GHz, a 500 GB SSD, and 16 GB RAM. The programming language used is Python, along with standard libraries including Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, Keras, Scikit-learn, and others. It should be further noted here that, in our experimental assessment and analysis, we only explore the Bangla language as the input, so as to better test the sensitivity of the results and the independence of the core technique, similarly to other studies in the field (e.g., [46]). As a comparison, we considered several state-of-the-art ML algorithms (see Section 1), namely DT, RF, LR, and MLP.

5.2. Performance Evaluation Metrics

The parameters used to evaluate the model's performance are (see Section 1) as follows: accuracy, precision, recall, F1-score, MAE, MSE, RMSE, confusion matrices, and ROC curves. While accuracy, precision, recall, and F1-score are classical ML metrics [32], the other MAE, MSE, and RMSE parameters are defined as follows:

$$\boldsymbol{MAE} = \frac{\sum_{i=1}^{n} \boldsymbol{predicted}(\boldsymbol{i}) - \boldsymbol{actual}(\boldsymbol{i})}{\boldsymbol{n}}, \tag{2}$$

$$\boldsymbol{MSE} = \frac{\sum_{i=1}^{n} (\boldsymbol{predicted}(\boldsymbol{i}) - \boldsymbol{actual}(\boldsymbol{i}))^2}{\boldsymbol{n}}, \tag{3}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\boldsymbol{predicted}(\boldsymbol{i}) - \boldsymbol{actual}(\boldsymbol{i}))^2}{\boldsymbol{n}}} \tag{4}$$

where n is the total number of samples; $actual(n)$ is the set of actual values with n samples; $predicted(n)$ is the set of predicted values with n samples.

Confusion matrices are defined as a tool used to evaluate the performance of a classification model on a test dataset. They are particularly useful when working with supervised classification problems, when one is trying to predict categories or classes for observations based on certain characteristics.

ROC curves are a graph used to evaluate the performance of a binary classifier, such as an ML model that must distinguish between two classes, for example, positive and negative. ROC curves display the relationship between the true positive rate (TPR) and the false positive rate (FPR) as the classifier's decision threshold varies. The TPR represents the fraction of positive examples correctly classified as positive, while the FPR represents the fraction of negative examples incorrectly classified as positive. The ROC curve illustrates

the trade-off between the sensitivity (TPR) and specificity ($1 - \text{FPR}$) of the classifier as the threshold varies. An ideal classifier will have an area under the curve (AUC) equal to 1, meaning that it can perfectly distinguish between classes. A random classifier will have an AUC of 0.5, as its discriminatory ability is akin to tossing a coin. The higher the AUC, the better the classifier's performance in distinguishing between classes.

K-fold cross-validation (CV) [32] is a common technique used in ML to assess the performance of a predictive model while simultaneously reducing the risk of overfitting. This technique divides the available dataset into *K* parts, known as folds, of approximately equal size. Subsequently, the model is trained and evaluated *K* times, each time using a different part of the dataset as the test set and the remaining $K - 1$ parts as the training set. The goal of *K*-fold cross-validation is to evaluate how well the model generalizes to unseen data, avoiding the risk of misleading evaluations due to specific random divisions of the dataset. In our model, *K* has been set to 10.

5.3. Analysis of Experimental Results

Experiments for classification were conducted using both the binary method and the multi-class method. In the following subsection, we present the results related to both methods.

Binary results analysis. The performance analysis of the binary classification is displayed in Figures 3 and 4.

The confusion matrices are depicted in Figure 5.

The ROC curves are presented in Figure 6.

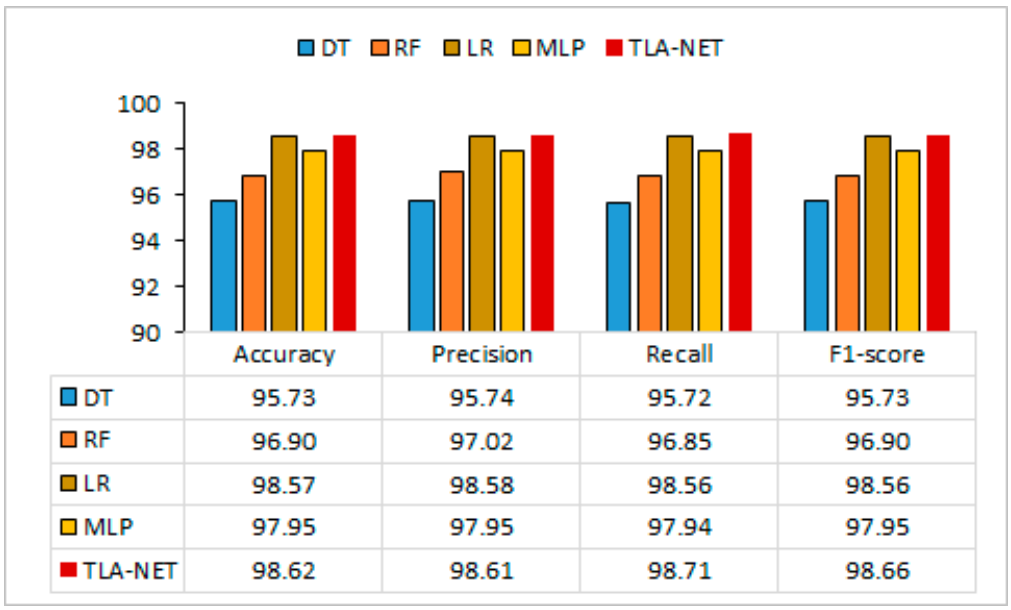
From Figures 3–6, in the analysis of binary results, we find that TLA-NET produces a better outcome in terms of both performance rate and a lower error rate, as well as a better AUC score.

Multi-class results analysis. The performance analysis of the multi-class classification is displayed in Figures 7 and 8.

The confusion matrices are depicted in Figure 9.

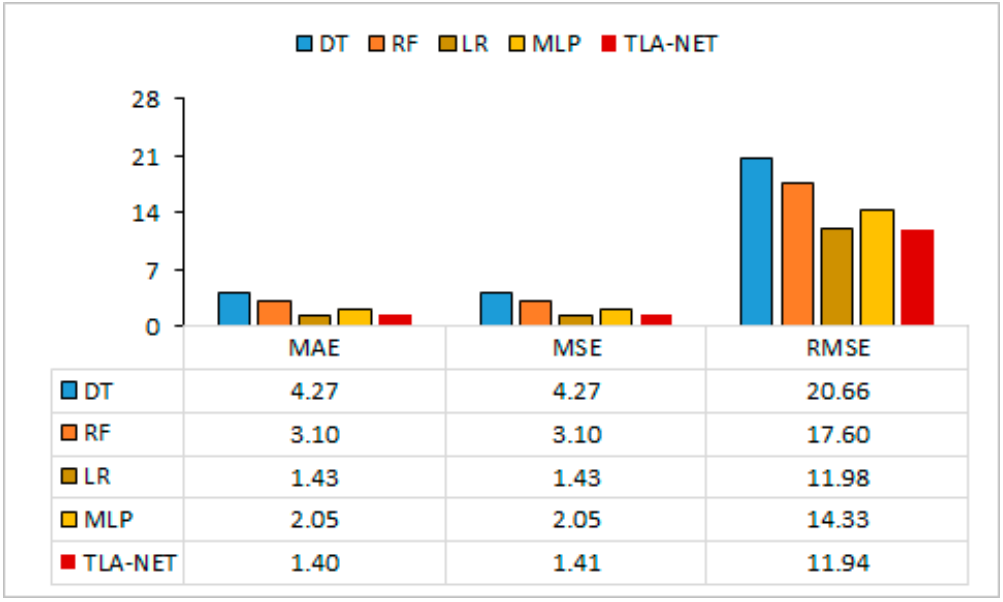
The ROC curves are presented in Figure 10.

From Figures 7–10, in the analysis of the multi-class results, we find that TLA-NET yields better performance rates and a lower error rate, as well as a superior AUC score.



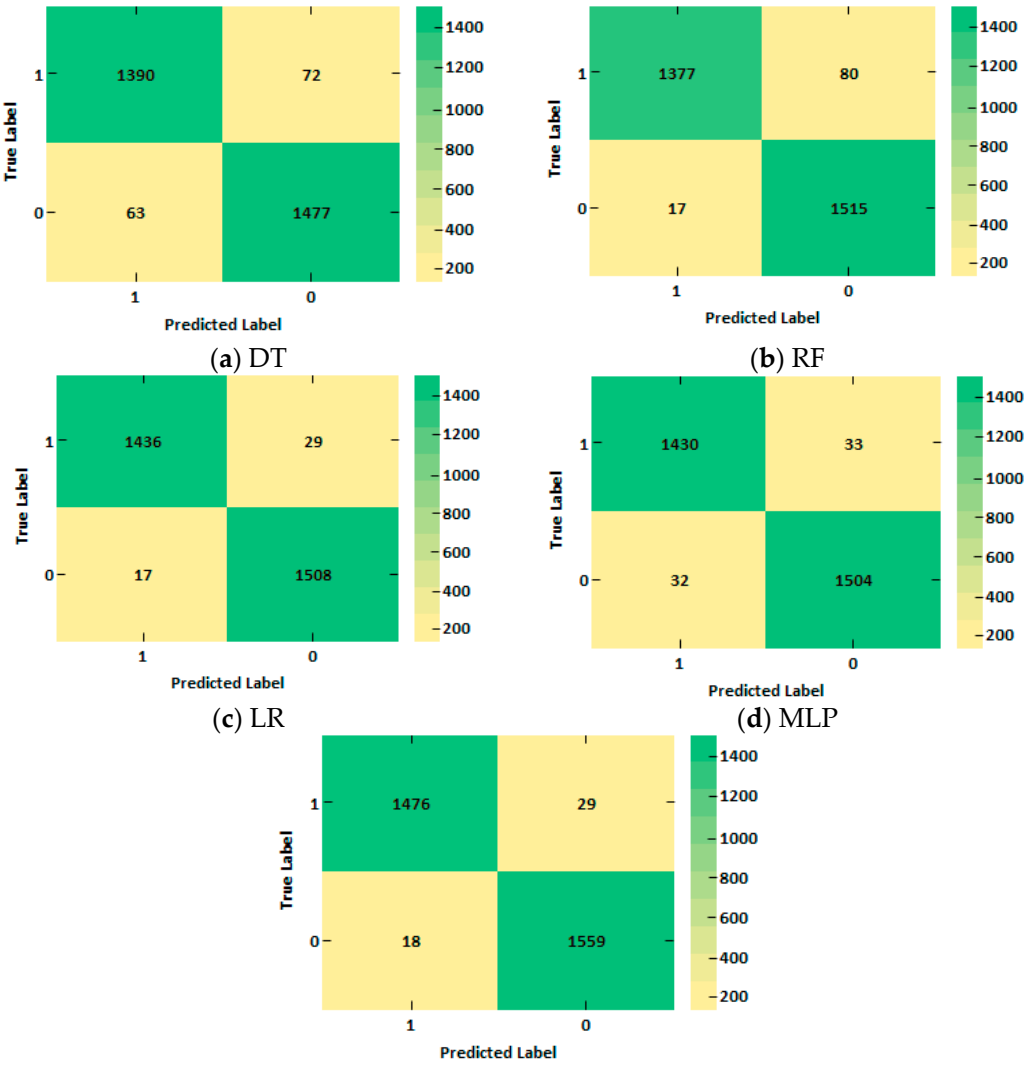
	Accuracy	Precision	Recall	F1-score
DT	95.73	95.74	95.72	95.73
RF	96.90	97.02	96.85	96.90
LR	98.57	98.58	98.56	98.56
MLP	97.95	97.95	97.94	97.95
TLA-NET	98.62	98.61	98.71	98.66

Figure 3. Binary classification performance evaluation of different models with the proposed approach.



	MAE	MSE	RMSE
DT	4.27	4.27	20.66
RF	3.10	3.10	17.60
LR	1.43	1.43	11.98
MLP	2.05	2.05	14.33
TLA-NET	1.40	1.41	11.94

Figure 4. Binary classification error evaluation of different models with the proposed approach.



(**a**) DT

(**b**) RF

(**c**) LR

(**d**) MLP

(**e**) TLA-NET

Figure 5. Confusion matrices for binary classification labels across different models.

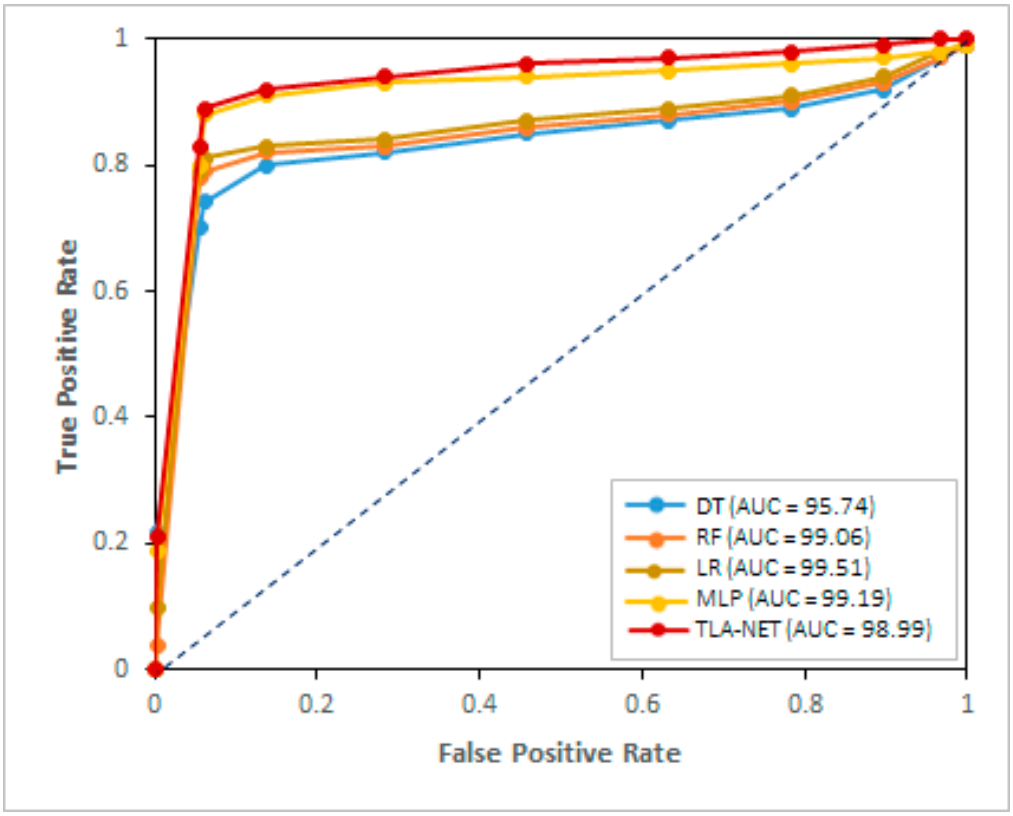
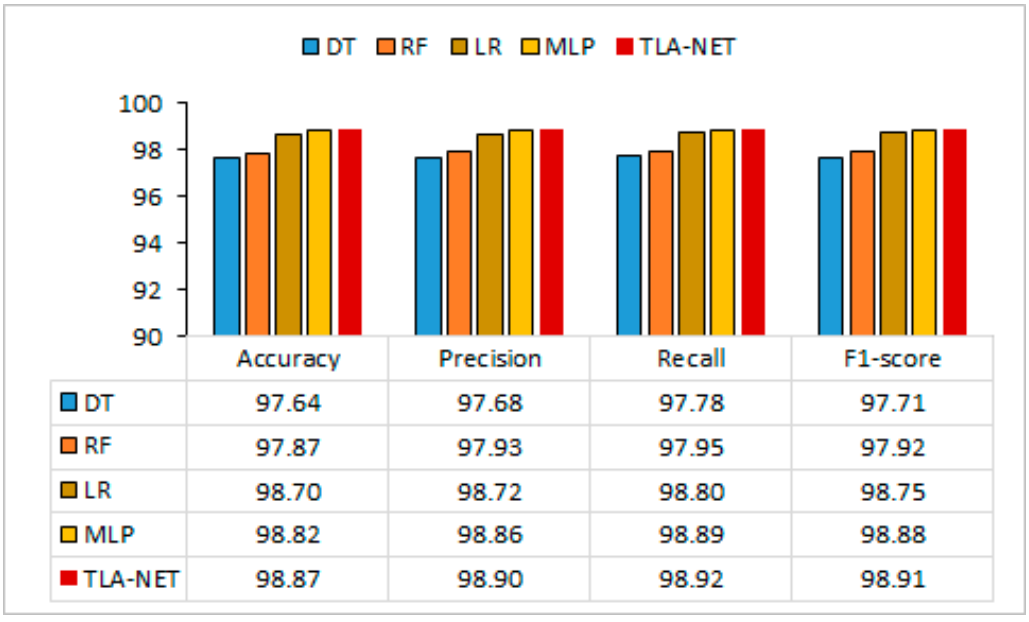
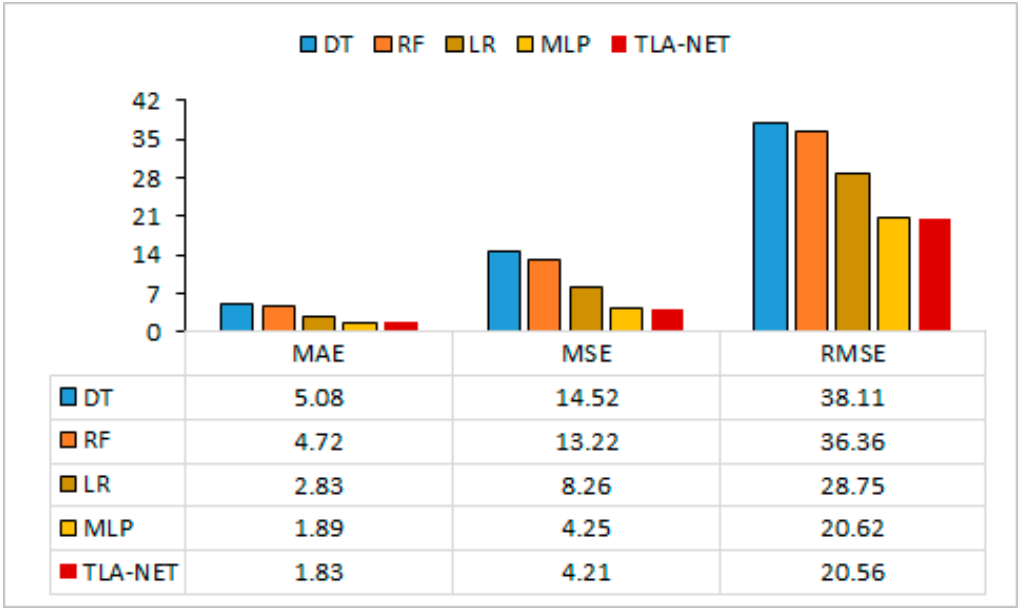


Figure 6. ROC curves for the binary classification labels across different models.



	Accuracy	Precision	Recall	F1-score
DT	97.64	97.68	97.78	97.71
RF	97.87	97.93	97.95	97.92
LR	98.70	98.72	98.80	98.75
MLP	98.82	98.86	98.89	98.88
TLA-NET	98.87	98.90	98.92	98.91

Figure 7. Multi-class classification performance evaluation of different models with the proposed approach.



	MAE	MSE	RMSE
DT	5.08	14.52	38.11
RF	4.72	13.22	36.36
LR	2.83	8.26	28.75
MLP	1.89	4.25	20.62
TLA-NET	1.83	4.21	20.56

Figure 8. Multi-class classification error evaluation of different models with the proposed approach.

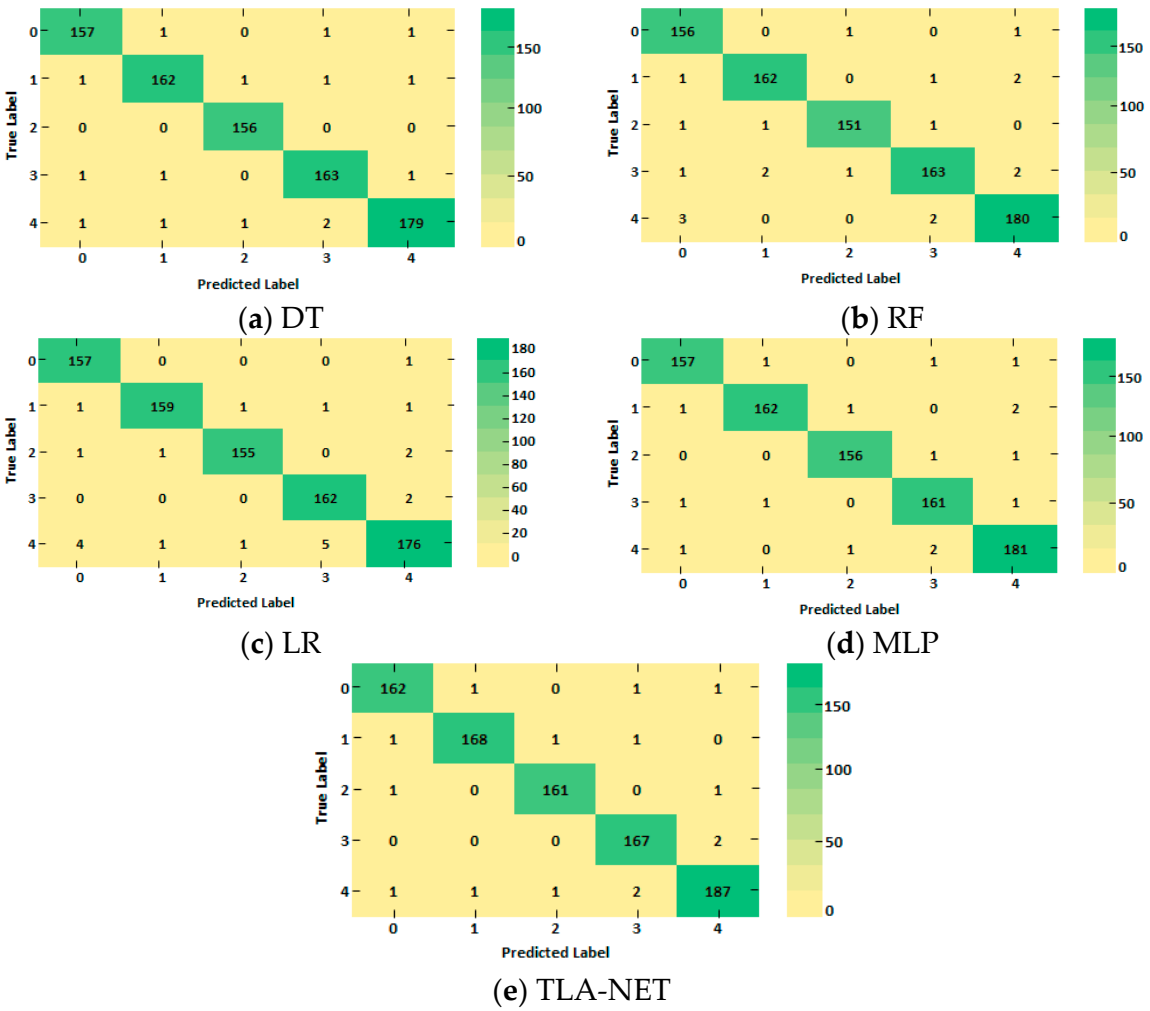


Figure 9. Confusion matrices for multi-class classification labels across different models.

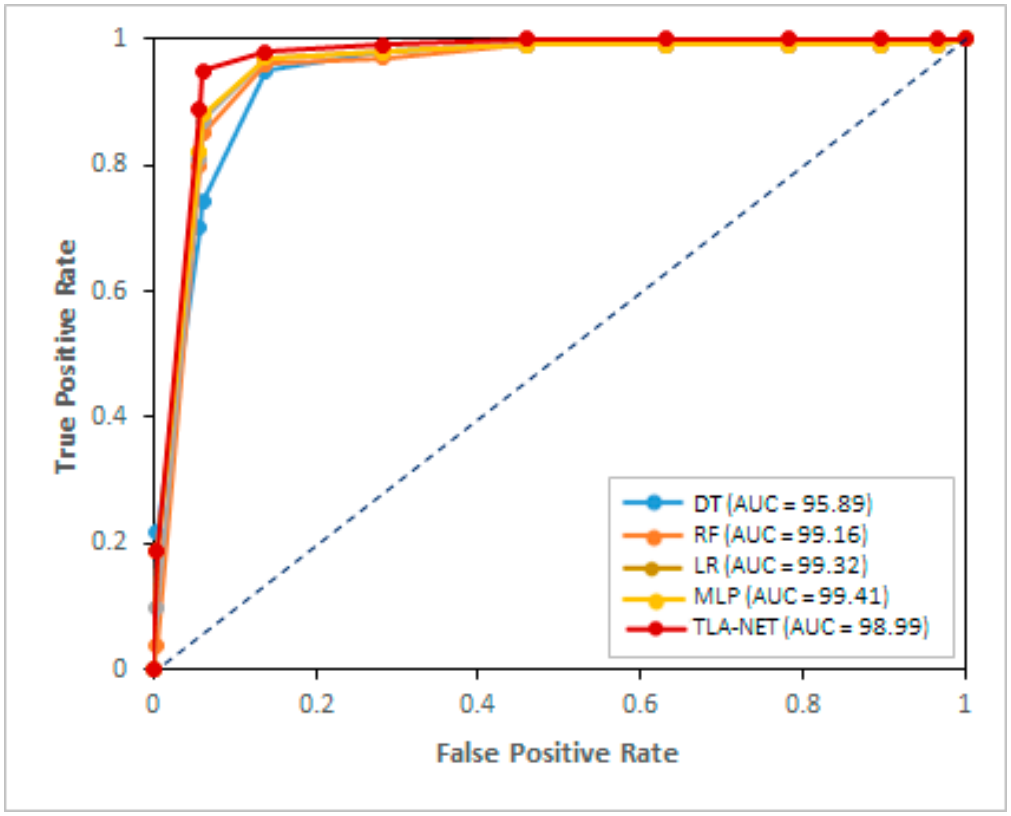


Figure 10. ROC curves for multi-class classification labels across different models.

6. Conclusions and Future Work

Bullying often occurs on social media because it provides aggressors with a place to hide, making them difficult to track down and avoiding direct confrontation. This situation makes it challenging to identify instances of bullying, resulting in suffering for the victims and often causing mental distress, sadness, and, in severe cases, even suicide. In response to this toxic climate on social media, it is essential to take proactive as well as reactive action. With the goal of classifying cyberbullying, this study has presented the process of automatically translated data generation and how deep learning models perform on this dataset. Furthermore, a modified `LSTM-autoencoder` model has been developed, which offers highly accurate performance on both noise-free and noisy data. The raw data represent the dataset we collected from the organization. The semi-noisy dataset refers to the augmented data we added to the raw data, while the noisy dataset refers to the data we translated from the Bangla and Hindi languages to the English language dataset. We have conducted a comparative analysis of our proposed model against some state-of-the-art `ML` approaches. To evaluate the model performances, we have used several state-of-the-art metrics and composite evaluation tools.

Our performance evaluation has shown that, in comparison to `ML` models, the proposed `TLA-NET` model achieved the best accuracy performance on the raw Bangla dataset. Overall, the results obtained from traditional models are not satisfactory. However, our developed model has overcome the shortcomings of existing models and has achieved excellent accuracy performance, as described. Using our approach, future researchers will be able to analyze various issues associated with text datasets that were previously neglected due to data availability. With our `TLA-NET` model, they will be able to detect cyberbullying with high precision on social media.

Following the results of our study, we can derive several implications for future research efforts, which could signficantly extend our efforts.

Anomaly detection. When dealing with cyberbullying comments, a relevant problem relates to how to detect anomalies in text, perhaps following some state-of-the-art proposals (e.g., [47]). This will increase the quality and the accuracy of both the target dataset and the ML analysis performed on top of it.

Visualization metaphors. A relevant future research direction for cyberbullying detection and analysis is represented by the issue of supporting advanced visualization metaphors (e.g., [48]) in order to improve data visualization paradigms over cyberbullying activities (e.g., [49]). The latter property will improve the precision of the analysis and the effectiveness of actions for combatting cyberbullying in society.

Performance and scalability. Last but not least, the performance and scalability aspects of algorithms for supporting cyberbullying detection and analysis play a relevant role, like in actual big data trends (e.g., [50–52]). In fact, social media can easily originate datasets with classical 3V characteristics (volume, velocity, and variety), so specialized solutions are necessary to ensure executions that scale with the data growth.

Author Contributions: Conceptualization, A.C. and M.S.A.; Methodology, A.C. and M.S.A.; Validation, H.S. and P.G.B.; Data curation, M.S.A.; Writing—original draft, M.S.A.; Writing—review & editing, A.C.; Supervision, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union-NextGenerationEU.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are publicly available.

Conflicts of Interest: The authors declare that they have no conflicts of interest. They also declare that they do not have any competing financial interests.

References

1. Abarna, S.; Sheeba, J.I.; Jayasrilakshmi, S.; Devaneyan, S.P. Identification of Cyber Harassment and Intention of Target Users on Social Media Platforms. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105283. [CrossRef] [PubMed]
2. Giumetti, G.W.; Kowalski, R.M. Cyberbullying via Social Media and Well-being. *Curr. Opin. Psychol.* **2022**, *45*, 101314. [CrossRef] [PubMed]
3. Akram, W.; Kumar, R. A Study on Positive and Negative Effects of Social Media on Society. *Int. J. Comput. Sci. Eng.* **2017**, *5*, 351–354. [CrossRef]
4. Trotter, J.L.; Allen, N.E. The Good, the Bad, and the Ugly: Domestic Violence Survivors' Experiences with their Informal Social Networks. *Am. J. Community Psychol.* **2009**, *43*, 221–231. [CrossRef] [PubMed]
5. Rosa, H.; Pereira, N.S.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic Cyberbullying Detection: A Systematic Review. *Comput. Hum. Behav.* **2019**, *93*, 333–345. [CrossRef]
6. Nandhini, B.; Sheeba, J. Online Social Network Bullying Detection Using Intelligence Techniques. *Procedia Comput. Sci.* **2015**, *45*, 485–492. [CrossRef]
7. Hemphill, S.A.; Kotevski, A.; Heerde, J.A. Longitudinal Associations Between Cyberbullying Perpetration and Victimization and Problem Behavior and Mental Health Problems in Young Australians. *Int. J. Public Health* **2015**, *60*, 227–237. [CrossRef]
8. Ioannou, A.; Blackburn, J.; Siringhini, G.; De Chrisiofaro, E.; Kouriellis, N.; Sirivianos, M.; Zaphiris, P. From Risk Factors to Detection and Intervention: A Metareview and Practical Proposal for Research on Cyberbullying. In Proceedings of the 2017 IEEE IST-Africa Week Conference, Windhoek, Namibia, 30 May–2 June 2017; pp. 1–8.
9. Al Mazari, A. Cyber-Bullying Taxonomies: Definition, Forms, Consequences and Mitigation Strategies. In Proceedings of the 5th IEEE International Conference on Computer Science and Information Technology, Amman, Jordan, 27–28 March 2013; pp. 126–133.
10. Perera, A.; Fernando, P. Accurate Cyberbullying Detection and Prevention on Social Media. *Procedia Comput. Sci.* **2021**, *181*, 605–611. [CrossRef]
11. Evangelio, C.; Rodriguez-Gonzalez, P.; Fernandez-Rio, J.; Gonzalez-Villora, S. Cyberbullying in Elementary and Middle School Students: A Systematic Review. *Comput. Educ.* **2022**, *176*, 104356. [CrossRef]

12. Schneider, S.K.; O'Donnell, L.; Stueve, A.; Coulter, R.W. Cyberbullying, School Bullying, and Psychological Distress: A Regional Census of High School Students. *Am. J. Public Health* **2012**, *102*, 171–177. [CrossRef]
13. Sharif, O.; Hoque, M.M. Tackling Cyber-Aggression: Identification and Fine-Grained Categorization of Aggressive Texts on Social Media Using Weighted Ensemble of Transformers. *Neurocomputing* **2022**, *490*, 462–481. [CrossRef]
14. Isaza, G.; Munõz, F.; Castillo, L.; Buitrago, F. Classifying Cybergrooming for Child Online Protection Using Hybrid Machine Learning Model. *Neurocomputing* **2022**, *484*, 250–259. [CrossRef]
15. Ramana, D.; Reddy, T.H. Detection of Online Hate in Social Media Platforms for Twitter Data: A Prefatory Step. In Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Aizawl, India, 25–26 June 2021; pp. 411–419.
16. Roy, P.K.; Bhawal, S.; Subalalitha, C.N. Hate Speech and Offensive Language Detection in Dravidian Languages Using Deep Ensemble Framework. *Comput. Speech Lang.* **2022**, *75*, 101386. [CrossRef]
17. Anand, M.; Sahay, K.B.; Ahmed, M.A.; Sultan, D.; Chandan, R.R.; Singh, B. Deep Learning and Natural Language Processing in Computation for Offensive Language Detection in Online Social Networks by Feature Selection and Ensemble Classification Techniques. *Theor. Comput. Sci.* **2022**, *943*, 203–218. [CrossRef]
18. Hossain, M.A.; Quaddus, M.; Warren, M.; Akter, S.; Pappas, I. Are You a Cyberbully On Social Media? Exploring the Personality Traits Using a Fuzzy-set Configurational Approach. *Int. J. Inf. Manag.* **2022**, *66*, 102537. [CrossRef]
19. Bennett, K.P.; Colin, C. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explor.* **2000**, *2*, 1–13. [CrossRef]
20. Alotaibi, M.; Alotaibi, B.; Razaque, A. A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media. *Electronics* **2021**, *10*, 2664. [CrossRef]
21. Ahmed, M.F.; Mahmud, Z.; Biash, Z.T.; Ryen, A.A.N.; Hossain, A.; Ashraf, F.B. Cyberbullying Detection Using Deep Neural Network From Social Media Comments in Bangla Language. *arXiv* **2021**, arXiv:2106.04506.
22. Kumari, K.; Singh, J.P. AI_ML_NIT_Patna @ TRAC-2: Deep Learning Approach for Multi-lingual Aggression Identification. In Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 21–23 May 2020; pp. 113–119.
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
24. Lagler, K.; Schindelegger, M.; Bôhm, J.; Krásná, H.; Nilsson, T. GPT2: Empirical Slant Delay Model for Radio Space Geodetic Techniques. *Geophys. Res. Lett.* **2013**, *40*, 1069–1073. [CrossRef]
25. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
26. Graves, A. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
27. Nguyen, H.; Tran, K.P.; Thomassey, S.; Hamad, M. Forecasting and Anomaly Detection Approaches Using LSTM and LSTM Autoencoder Techniques with the Applications in Supply Chain Management. *Int. J. Inf. Manag.* **2021**, *57*, 102282. [CrossRef]
28. Hou, B.; Yang, J.; Wang, P.; Yan, R. LSTM-based Auto-encoder Model for ECG Arrhythmias Classification. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 1232–1240. [CrossRef]
29. Said Elsayed, M.; Le-Khac, N.A.; Dev, S.; Jurcut, A.D. Network Anomaly Detection Using LSTM Based Autoencoder. In Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks, Alicante, Spain, 16–20 November 2020; pp. 37–45.
30. Church, K.W. Word2vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162. [CrossRef]
31. Ahmed, M.F.; Mahmud, Z.; Biash, Z.T.; Ryen, A.A.N.; Hossain, A.; Ashraf, F.B. Bangla Online Comments Dataset. *Mendeley Data, Version 1*, 2021. [CrossRef]
32. Zhou, Z.H. *Machine Learning*; Springer Nature: Berlin/Heidelberg, Germany, 2021.
33. Rokach, L.; Maimon, O. Decision Trees. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2005; pp. 165–192.
34. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227. [CrossRef]
35. LaValley, M.P. Logistic Regression. *Circulation* **2008**, *117*, 2395–2399. [CrossRef]
36. Almeida, L.B. Multilayer Perceptrons. In *Handbook of Neural Computation*; CRC Press: Boca Raton, FL, USA, 2020; chapters 1–2.
37. Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access* **2019**, *7*, 70701–70718. [CrossRef]
38. Shetty, N.P.; Muniyal, B.; Priyanshu, A.; Das, V.R. FedBully: A Cross-Device Federated Approach for Privacy Enabled Cyber Bullying Detection using Sentence Encoders. *J. Cyber Secur. Mobil.* **2023**, *12*, 465–496. [CrossRef]
39. Dong, Y.; Li, Y.; Zhao, D.; Shen, G.; Zeng, Y. Bullying10K: A Large-Scale Neuromorphic Dataset towards Privacy-Preserving Bullying Recognition. In Proceedings of the Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.

40. Akter, M.S.; Shahriar, H.; Cuzzocrea, A.; Wu, F.; Rodriguez-Cardenas, J. A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data. In Proceedings of the 2023 IEEE International Conference on Big Data, Sorrento, Italy, 15–18 December 2023; pp. 5418–5427.
41. Mali, M.K.; Pawar, R.R.; Shinde, S.A.; Kale, S.D.; Mulik, S.V.; Jagtap, A.A.; Tambewagh, P.A.; Rajput, P.U. Automatic Detection of Cyberbullying Behaviour on Social Media using Stacked Bi-Gru Attention with BERT Model. *Expert Syst. Appl.* **2025**, *262*, 125641. [CrossRef]
42. Al-Khasawneh, M.A.; Faheem, M.; Alarood, A.A.; Habibullah, S.; Alsolami, E.A. Toward Multi-Modal Approach for Identification and Detection of Cyberbullying in Social Networks. *IEEE Access* **2024**, *12*, 90158–90170. [CrossRef]
43. Jamjoom, A.A.; Karamti, H.; Umer, M.; Alsubai, S.; Kim, T.; Ashraf, I. RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features. *IEEE Access* **2024**, *12*, 58950–58959. [CrossRef]
44. Akhter, A.; Acharjee, U.K.; Talukder, M.A.; Islam, M.M.; Uddin, M.A. A Robust Hybrid Machine Learning Model for Bengali Cyber Bullying Detection in Social Media. *Nat. Lang. Process. J.* **2023**, *4*, 100027. [CrossRef]
45. Bhattacharya, S.; Singh, S.; Kumar, R.; Bansal, A.; Bhagat, A.; Dawer, Y.; Lahiri, B.; Ojha, A.K. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. *arXiv* **2020**, arXiv:2003.07428.
46. Bastiaensens, S.; Vandebosch, H.; Poels, K.; Van Cleemput, K.; DeSmet, A.; De Bourdeaudhuij, I. Cyberbullying on Social Network Sites. An Experimental Study into Bystanders' Behavioural Intentions to Help the Victim or Reinforce the Bully. *Comput. Hum. Behav.* **2014**, *31*, 259–271. [CrossRef]
47. Langone, R.; Cuzzocrea, A.; Skantzos, N. Interpretable Anomaly Prediction: Predicting Anomalous Behavior in Industry 4.0 Settings Via Regularized Logistic Regression Tools. *Data Knowl. Eng.* **2020**, *130*, 101850. [CrossRef]
48. Barkwell, K.E.; Cuzzocrea, A.; Leung, C.K.; Ocran, A.A.; Sanderson, J.M.; Stewart, J.A.; Wodi, B.H. Big Data Visualisation and Visual Analytics for Music Data Mining. In Proceedings of the 22nd IEEE International Conference Information Visualisation, Fisciano, Italy, 10–13 July 2018; pp. 235–240.
49. Srinath, A.S.; Johnson, H.; Dagher, G.G.; Long, M. BullyNet: Unmasking Cyberbullies on Social Networks. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 332–344. [CrossRef]
50. Yu, B.; Cuzzocrea, A.; Jeong, D.H.; Maydebura, S. On Managing Very Large Sensor-Network Data Using Bigtable. In Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Ottawa, ON, Canada, 13–16 May 2012; pp. 918–922.
51. Setiawan, Y.; Maulidevi, N.U.; Surendro, K. The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification. *Data Sci. J.* **2024**, *23*, 31. [CrossRef]
52. Singh, N.M.; Sharma, S.K. An Efficient Automated Multi-Modal Cyberbullying Detection using Decision Fusion Classifier on Social Media Platforms. *Multimed. Tools Appl.* **2024**, *83*, 20507–20535. [CrossRef]