Research Article

Fawaz Waselallah Alsaade* and Mohammed Saeed Alzahrani

Transformer learning-based neural network algorithms for identification and detection of electronic bullying in social media

https://doi.org/10.1515/dema-2023-0118
received March 27, 2023; accepted August 19, 2023

Abstract: The global phenomenon known as cyberbullying is a form of modern harassment that cannot be entirely stopped but can be avoided. Most current solutions to the cyberbullying problem have relied on tools and methods to identify online bullying. However, end users do not have free access to these tools. The goal of this study is to create a model to combat cyberbullying on social media sites based on users' appearance. In this article, we present a cyberbullying detection system constructed using the Word2Vec word-embedding method and a deep learning convolutional neural network combined with bidirectional long short-term memory (CNN-BiLSTM), as well as the XLM-Roberta transformer, to develop a model for cyberbullying detection. We carried out two experiments based on binary (hate speech or non-hate speech bullying comments) and multiclass (religion, age, gender, ethnicity, and non-bullying tweets) datasets collected from Kaggle online discussions and Twitter. To evaluate the model's performance, we used standard measurement metrics, such as precision, recall, *F*1-score, and accuracy. Through a comparison of the results, it is noted that the XLM-Roberta model outperformed the CNN-BiLSTM model, resulting in 84% accuracy using the Kaggle online discussion dataset and 94% accuracy using the Twitter dataset.

Keywords: artificial intelligence, cyberbullying, deep learning, social media

MSC 2020: 68T50, 68T07

1 Introduction

Most people around the world today utilize social media platforms daily for communication, and because these platforms are so pervasive and provide users with a significant level of anonymity, it is possible for anybody, at any time, and in any location to be the target of cyberbullying. The United Nations Children's Fund issued a notification on April 15, 2020, in response to the increased risk of harassment and bullying during the coronavirus disease of 2019 (COVID-19) pandemic caused by widespread school closures, rising screen time, and declining in-person social connections. The notification was published in response to the increased risk of harassment and bullying during the pandemic. Cyber harassment is defined by Englander et al. [1] as the use of digital information to intimidate an individual or group of individuals online, typically by delivering messages that are generally frightening or threatening. Meanwhile, cyberbullying is linked to physical bullying and is currently being studied. The statistics on online bullying are dire. In middle and high school, 36.5% of students report experiencing online harassment and threats, while 87% report witnessing bullying in action. The effects

*** Corresponding author: Fawaz Waselallah Alsaade,** College of Computer Science and Information Technology, King Faisal University, P.O. Box 4000 Al-Ahsa, Saudi Arabia, e-mail: falsaade@kfu.edu.sa

Mohammed Saeed Alzahrani: College of Computer Science and Information Technology, King Faisal University, P.O. Box 4000 Al-Ahsa, Saudi Arabia, e-mail: malzahrani@kfu.edu.sa

of cyberbullying can include poor academic performance, dissatisfaction, and even suicide attempts. The three main strategies used today to stop online harassment are teaching "internet street smarts," looking for warning signs, and counseling [2]. Even though in the United States, it is prohibited to engage in cyberbullying in all 50 states, most of the relevant laws do not apply to situations outside the classroom. Although social media platforms, such as Facebook, Twitter, Instagram, and Snapchat, provide information on online bullying, they do not provide practical methods for combating it. Considering that 90% of cyberbullying instances go unreported, having a reliable technique for detecting intelligent user-generated content is crucial. Even though many organizations work to increase awareness of cyberbullying, the frequency of incidents when someone is persistently harmed or bullied through digital technology is on the rise [3]. According to a study by Slonje and Smith [4], three billion people frequently visit social networking sites to connect with one another. Facebook and other social networks are undoubtedly useful, but they can also be abused. Online harassment is referred to as the malevolent use of digital technology to harm an individual or a group of individuals [5]. Cyberbullying, which may quickly reach a vast number of people, has more severe and pervasive effects than traditional bullying. Furthermore, it can be challenging or even impossible to delete harmful data from online resources. Cyberbullying has been implicated in such mental health problems as despair, low self-esteem, weariness, and even suicidal tendencies [6], despite the lack of evidence that it causes physical harm to the victim. Over the past 10 years, cyberbullying has become more common, particularly among children and teenagers. According to a recent study [7], 37, 26, 26, and 20% of children in India, the United States, South Africa, and Turkey, respectively, were victims of cyberbullying in 2018. According to these data, the severity of this problem is increasing at a quick rate and is unrelated to the degree of economic growth in nation. In Sweden, one of the most industrialized nations in the world, there was an increase in incidents of cyberbullying between the years 2011 and 2018. As a direct result of this study, there has been an increase in awareness of cyberbullying in many countries, according to numerous studies conducted using machine learning approaches to detect online bullying. Most research on this topic is in English and has made use of text mining techniques, which are similar to those applied in sentiment analysis studies. Social media posts should not be considered independent texts because they are dynamic and context dependent in nature. To recognize and forecast various content containing cyberbullying-related behavioral patterns posted on social media platforms, such as Twitter and Kaggle online discussions, we intend to construct a hybrid deep learning model that includes a combination of a convolutional neural network and bidirectional long short-term memory (CNN-BiLSTM) for the detection and classification of tasks. The contents in question contain offensive and hateful speech and are related to religion, age, gender, and racism. The contributions of this research are as follows:

- Covering and analyzing the most common types of online cyberbullying on social media platforms.
- Conducting binary and multiclass experiments for online cyberbullying detection.
- Designing a hybrid-structure deep learning model.
- Comparing the performance of state-of-the-art transformers, such as the XLM-Roberta, with the combined CNN-BiLSTM model to generate more reliable classification results.

2 Background of study

This section discusses the existing relevant literature. Based on the supervised machine learning technique, as well as on the existing body of research regarding the phenomenon of cyberbullying, the first investigation into the automated identification of cyberbullying online was conducted by Yin et al. [8]. The authors analyzed three data sets in order to detect cyberbullying on three distinct websites. The first two datasets were amassed from various social networking sites like Reddit, while the third was produced using the video game portal Kongregate. The classification procedure used a linear kernel classifier in addition to feature extraction strategies (N-grams and term frequency-inverse term frequency [TF-IDF]). While the study's experimental results were muddled, it laid the groundwork for future studies. Text answers from the Formspring.me website were analyzed by Reynolds et al. [9]. In their capacity as classifiers, these individuals used techniques such as C4.5, k-nearest neighbor (KNN), and support vector machines (SVM) to the data. Experimental results demonstrated that the C4.5 decision tree classifier, with a reported detection accuracy of 78.5%, beat both the KNN and SVM classifiers. Kumar et al. [10] applied soft computing approaches to detect cyberbullying, notably via social media platforms. The next phase was applying the multiclass classifier method to the text in order to categorize it (e.g., intellectual ability, sexual orientation). These techniques were used to evaluate 4,500 comments extracted from Kaggle online discussions, and the classification accuracy was 80% [11]. In a study of gender-based cyberbullying, Dadvar et al. [12] suggested a methodology based on gender to identify harassment. They used two different vocabulary sets and found that this approach slightly improved the accuracy of machine learning classifiers [12]. After early investigations into cyberbullying detection, several studies were published that used a variety of methodologies. Kontostathis et al. [13] classified the most used terms most often while engaging in cyberbullying from a collection of messages sent over the Formspring.me platform. They created a model based on the essential dimensions of latent semantic indexing. According to the authors, the experimental results had an overall accuracy of 91.25%. Ptaszynski et al. [14] used learning classifiers and brute force search methods to identify patterns related to online cyberbullying. In their study, the categorization procedure utilized patterns retrieved from phrases, a method said to outperform previous cyberbullying-detection approaches, according to the database compiled by the Human Rights Center. Furthermore, Fiebrink and Gillies [15] argued that when studying machine learning from a human-centered perspective, it is important to explicitly acknowledge both human activity and the environments in which machine learning is applied. Consequently, human-centered optics can help us to carefully examine the integration of various stakeholders' viewpoints into the formation of a ground truth when training and testing cyberbullying detection algorithms. This is because, given its subjectivity [16] and its profoundly different health and social impacts on victims [17], cyberbullying has both objective and subjective effects. Regardless of whether the abuse was expected, the victim may interpret the experience in a completely different way from how the attacker intended. Moreover, human experts who analyze machine learning models, including clinicians and mental health specialists in the context of bullying detection, can assist in reducing the gaps between the models' performances and their social applications [18].

Ozel et al. [19] conducted research utilizing the Turkish language to study the identification online bullying. Streaming data from Twitter was utilized to build an evaluative dataset for the experimental work that was carried out. Using the bag-of-words methodology, a vector was assigned to each tweet, and the tweets were then classified using a variety of machine learning algorithms (SVM, naive Bayes [NB], C4.5, and KNN) to determine whether the tweets constituted mistreatment. In aspects of F-measure, NB significantly outperformed the other classification techniques, with a 79% accuracy rate. Most datasets examined for cyberbullying identification contain a sizable combination of features that could be employed to train machine learning models, such as data about gender, age, and many other characteristics from user profiles [20]. User-posted content on social media platforms also contains significant information and offers a variety of alternative representations of the data, making it useful for training classification models. The human decision-making processes that led to selecting these attributes were explored, as well as the degree to which preexisting models of cyberbullying were considered. Textual features were consistently regarded as essential to training the models in the articles reviewed here, because their datasets constituted textual comments. Word embeddings and sentiments were the types of textual information employed as features [21–26].

3 Materials and methods

This section outlines the main components of the planned cyberbullying detection system (CDS), which is used to examine and identify cyberbullying behavior on various websites, including Twitter and Kaggle online discussions. The steps used to build this framework are illustrated in Figure 1 and are described in the following subsections.

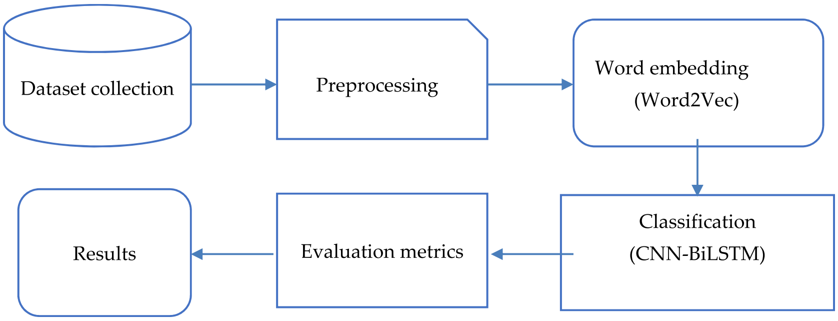

Figure 1: Proposed system based on AI approach to detect cyberbullying.

3.1 Dataset collection

In the suggested methodology, data collection is the most crucial step. We used two independent social media datasets, both gathered from the Kaggle platform, to enable the detection and analysis of various forms of online cyberbullying.

3.1.1 Cyberbullying multiclass dataset

This publicly available dataset is compiled from Twitter, a social networking site where users exchange and interact through short texts known as tweets. This dataset includes 39,869 tweet samples, which are divided into five categories: non-cyberbullying, gender, religion, age, and ethnicity. The non-cyberbullying class had 7,945 comments, the religion class 7,998 posts, the age class 7,992 posts, the gender class 7,973 posts, and the ethnicity class 7,961 posts.

3.1.2 Cyberbullying binary class dataset

Cyberbullying, such as hate speech, takes place online, and cyber harassment is often referred to as online bullying. This dataset, which includes 8,799 comment samples divided into 5,993 non-hate speech and 2,806 hate speech text comments, was acquired from Kaggle online discussions.

3.2 Preprocessing

Before using the modeling and transforming methods, preprocessing is used to clean and eliminate noise from the dataset. The purpose of text processing is to convey and modify social media post information such that it can be examined and categorized by the applied deep learning CNN-BiLSTM model. However, the datasets were cleaned using the following procedures:

- Having the contents of the social media posts used in the datasets cleaned up by getting rid of any extraneous words, emojis, spaces, or digits.
- Disassembling a sentence into its component words and other pieces (tokenization).
- Removing punctuation marks (?, !, :, ;, ”, ’) to make social media posts look more professional.
- Removing stop words, such as “the,” “a,” “an,” and “in.”
- Converting all uppercase words into lowercase words.
- Because each social media post included in this investigation was categorized as either cyberbullying or not by means of the deep learning neural network approach, it is imperative that all text sequences contained within the dataset have equal real-valued vectors in accordance with the post-padding sequence method.

3.3 Word2Vec

The term "word embeddings" is used in text mining tasks to refer to the vector representations of words for sentences and document classification. These vector representations typically take the form of a real-valued vector that embeds the original meaning of the text. As a result, it is expected that words located close to one another in the vector space would have definitions comparable to one another. Word embeddings can be obtained by combining several computational linguistic and feature learning techniques. We used Word 2Vec [27], which is a Google algorithm that employs a two-layer neural network to map the words in the given text into real-valued vector representations.

3.4 CNN-BiLSTM model

During this research project, a CNN-BiLSTM model was utilized to develop a CDS with the potential to be utilized on many social networking platforms. The structure and layers of the CNN-BiLSTM are shown in Figure 2.

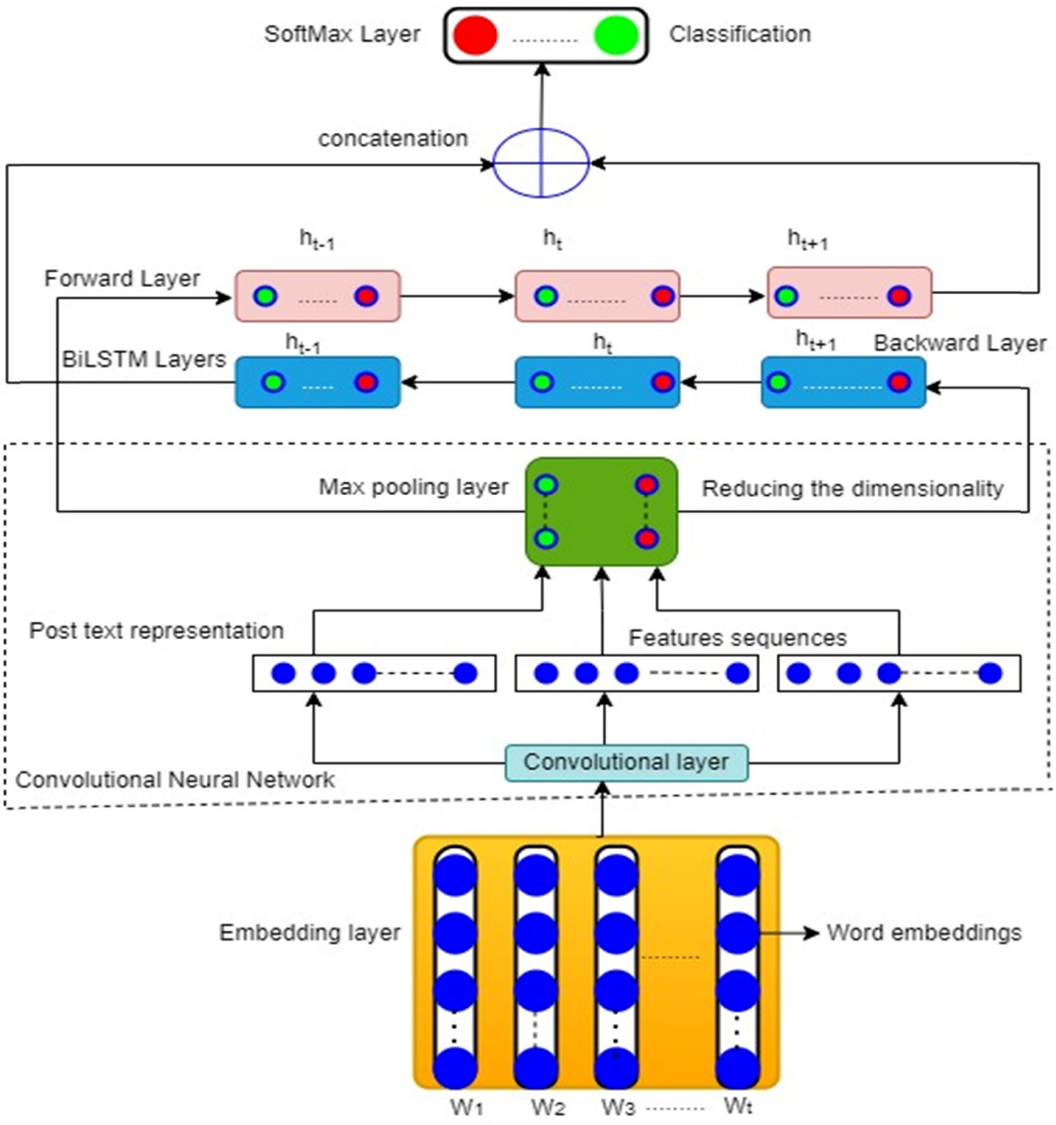

Figure 2: CNN-BiLSTM model.

3.4.1 Embedding layer

The embedding layer is the most employed neural network layer in natural language processing (NLP)-associated applications, such as language modeling [27–29]. In the CNN-BiLSTM model, the embedding layer was used to create an embedding matrix for each word in the training and testing sets. Maximum features, embedding dimensions, and input sequence length are the components of this layer. The 15,000 most frequently occurring words chosen from the training dataset are the maximum features (vocabulary size); they are shown in Figure 3 with the symbols W_1, W_2, …, W_n. Each word is vectorized into sequences of integer numbers according to the embedding word dimension, which was set to 100 dimensional vectors. The definition of the input sequence length is the average word length of each input social media post in the datasets, which was set to 98 words for the Twitter dataset and 162 words for the Kaggle online discussion dataset. These word lengths were chosen because they represent the types of content included in each dataset.

3.4.2 Convolutional layer

A CNN is a type of neural network used for computational intelligence. It is frequently used to find intricate patterns in NLP, computer vision, and image processing [30,31]. The architecture of a CNN is modeled on the visual cortex and closely resembles the way that neurons connect in the human brain. The mathematical operation function applied to an input data matrix that generates feature maps is what is meant to be understood when referring to the convolutional process. The convolution layer is the most important layer in building the CNN technique. It receives the input embedding matrix from an embedding layer and then uses said matrix to conduct mathematical operations. To extract sequence information and condense the input sequence's dimensions, filters are used to pass across the input embedding matrix. The equation for a convolutional operation is expressed as follows:

$$y_j^l = \sigma\left(\sum_{i=1}^{N_{i-1}} \mathrm{conv}(w_{i,j}^l, x_i^{l-1})\right) + b_j^l, \tag{1}$$

where N_{i-1} represents the number of feature maps, y_j^l represents the feature map of the word embeddings of the input sequences, $w_{i,j}^l$ stands for the convolutional kernel, b_j^l stands for the bias of the feature map, and represents the ReLU activation function for avoiding the overfitting problem in the conquered data.

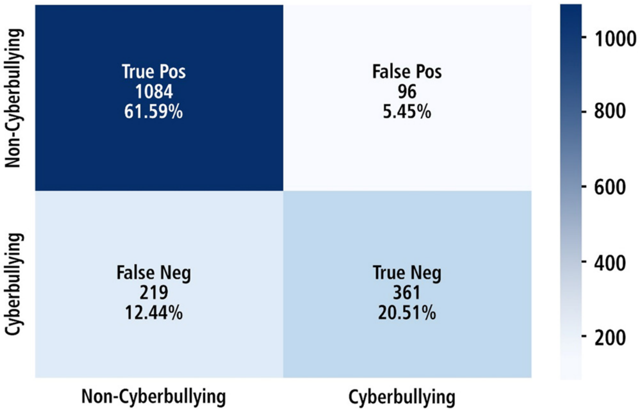

Figure 3: Confusion matrix for binary classification.

3.4.3 Max pooling layer

Using a max pooling layer to extract relevant global features from the feature map, we may minimize the training data's dimensionality and improve the classification accuracy. Max pooling layer equation.

$$Q_i = \mathrm{Max}(P_j^1, P_j^2, P_j^3, \ldots, P_j^t) \tag{2}$$

where Q_i specifies the output from the max pool and P_j^t illustrates the feature map before the maximization process.

3.4.4 BiLSTM layers

The long short-term memory (LSTM) network is a recurrent neural network model that may be used in AI and deep learning [32–34]. Some examples of these applications include sequence mining, text mining, NLP, and image processing. Memory cells that are actively involved in an LSTM can initially assign the results of prior knowledge concerning the input features to the output so that they can be matched to the subsequent input features. In addition, learning new features in an LSTM can only be done in one direction, and this direction is forward. Because of this, backward training is ignored, which leads to a decrease in the overall performance of the machine learning system. To address this issue, the BiLSTM was built with two independent hidden layers, each with its own configuration and a shared output. As can be seen in Figure 3, the characteristics of the data are learnt and processed in two layers that go in the opposite way. Input i_t, forget f_t, cell state c_t, and output o_t are the four gates that make up the LSTM. These gates have been defined using the following equations [34]:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \tag{3}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \tag{4}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \tag{5}$$

$$c_t = f_t c_{t-1} + i_t * \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c), \tag{6}$$

$$\overrightarrow{h_t} = o_t * \tanh(c_t), \tag{7}$$

$$\overleftarrow{h_t} = o_t * \tanh(c_t), \tag{8}$$

$$\tanh(x) = \frac{1 - e^{2x}}{1 - e^{2x}}, \tag{9}$$

$$H_t = (\overrightarrow{h_t} : \overleftarrow{h_t}), \tag{10}$$

where sig and tanh signify the sigmoid and tangent activation functions independently, x represents the input sequences, W and b represent the weight and bias factors, C_t is the cell state, h_t symbolizes the output of the LSTM cell, and H_t indicates the output of the bidirectional concatenation of the $\overrightarrow{h_t}$ forward and $\overleftarrow{h_t}$ backward LSTM layers at the current time t.

3.4.5 SoftMax layer

This layer is called the output layer in the CNN-BiLSTM model and is responsible for classifying the data sets' different outputs. This layer's neuron count may be found using the dataset's total class count as a reference. Two trials were conducted, one with a binary and the other with a multiclass cyberbullying detection dataset; therefore, two neurons and five neurons were independently configured for each experiment. Moreover, the activation function of the SoftMax layer computes the probability distribution of the input sequence vectors for each unique cyberbullying event and class in a studied dataset, such as age, gender, ethnicity, religion, and hate speech. In statistical terms, the SoftMax function is represented by the following equation:

$$\sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \tag{11}$$

where z represents the output layer neurons, and e is an exponential function serving as the non-linear function.

3.5 XLM-Roberta transformer approach

This subsection presents the detailed structure of the XLM-Roberta transformer-based model that was trained and tested for binary and multiclass cyberbullying social media post comments. This model was proposed by the Facebook Artificial Intelligence team for multilingual modeling tasks [35] and was trained on a large-scale vocabulary related to 100 different languages to perform various NLP tasks. One type of masked language model can learn semantic and contextual knowledge about words present in the input datasets by considering the advantages of XLM and RRoBERTa [36]. We used different layers and parameters to apply the model for the detection and classification of cyberbullying comments into various classes. Word embedding and the transformer encoder are the main two components of the XLM-Roberta model. In the fine-tuning step, the output feature vector C, which is equivalent to the initial identifier [CLS] in a single post comment, is reserved as the output of the model. Furthermore, the fully connected layer is linked to the word embedding layer, which consists of an attention mask, input Ids, and max-length parameters for inputting text sequences into the dataset, and at that time, the Softmax activation function is utilized to map the output results into various probability values corresponding to the number of dataset labels, which represent the various types of online cyberbullying.

3.6 Evaluation metrics

In this section, the evaluation criteria that were utilized to evaluate the performance of the applied hybrid deep learning model in its tests (the integrated CNN-BiLSTM model for classifying social media content into cyberbullying or non-cyberbullying posts). We used the model's confusion matrix to determine a number of standard performance metrics, including the number of false positives and negatives. The following equations provide definitions for the most often used performance metrics: precision, specificity, recall, accuracy, and $F1$ score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \times 100\%, \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \tag{14}$$

$$F1\text{-score} = 2 * \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \times 100\%. \tag{15}$$

These metrics are used for the model performance measurements, each of which has a range of values from 1 to 100, meaning with higher values of such metrics, the best results can be obtained.

4 Experimental results

This subsection investigates the findings of the experiments conducted using the hybrid deep learning model to identify and classify various types of cyberbullying, including age, religion, ethnicity, gender, hate speech, and non-bullying content. Two distinct real-world datasets were used to evaluate the proposed CDS with multiclass and binary classification settings. We split both dataset samples into training and validation sets before beginning training and optimizing the model, as indicated in Table 1.

Table 1: Cyberbullying dataset splitting

Dataset name	Total no. of samples	Training set 70%	Validation 10%	Testing 20%
Binary class Kaggle online discussion dataset	8,799	6,335	704	1,760
Multiclass Twitter dataset	39,869	28,705	3,987	7,176

4.1 Results of the binary classification

The proposed hybrid CNN-BiLSTM deep learning and XLM-Robert models were trained, validated, and tested with 70, 10, and 20% of the samples for an investigation into the effectiveness of binary bullying classification. In this scenario, the model is used to classify the binary class dataset into hate speech and non-hate speech bullying, and the results of the CNN-BiLSTM deep learning and XLM-Robert models are presented in Table 2. It is investigated whether the XLM-Robert model achieved a high accuracy of 90%.

Figure 3 shows the confusion matrices of the CNN-BiLSTM model for the binary classification dataset. The false-positive rate is 5.45%, and the true-positive rate is 61.59% for data classified as cyberbullying, and the true-negative rate is 20.51% for posts classified as not cyberbullying.

As observed from the confusion matrix of the XLM-Roberta model, depicted in Figure 4, of 1,760 text cyberbullying comments that constituted the testing set to authenticate the performance of the CNN-BiLSTM model for cyberbullying comment classification, 174 were obtained as misclassification rates.

Table 2: Online Kaggle debate dataset was used for testing the CNN-BiLSTM and XLM-Roberta models

Model Name	Results of the binary cyberbullying classification			
	Precision (%)	Recall (%)	*F*1-score (%)	Accuracy (%)
CNN-BiLSTM	79	62	70	82
XLM-Roberta	90	85.6	88	84

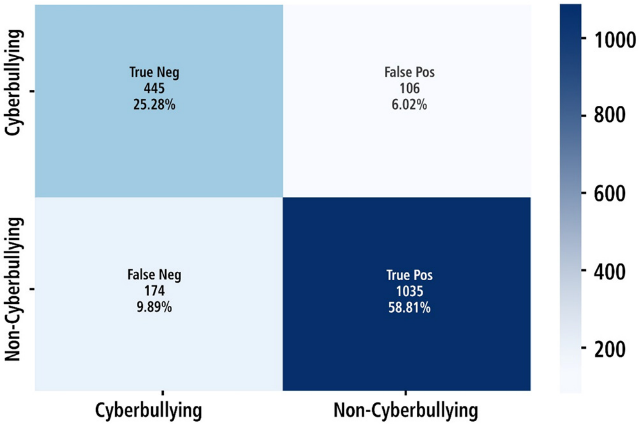

Figure 4: Confusion matrix of the XLM-Roberta model for binary classification.

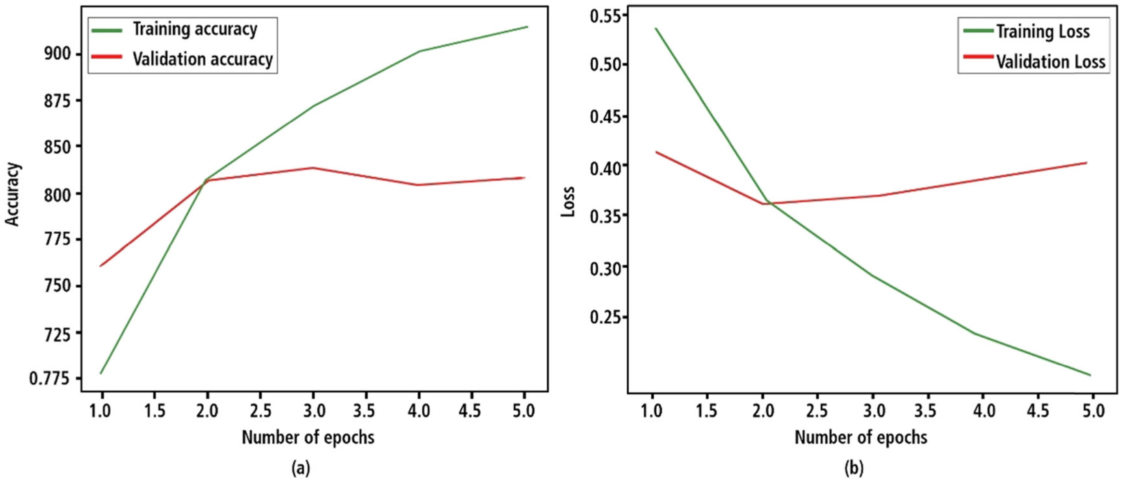

Figure 5: (a) Training accuracy and (b) loss of the CNN-BiLSTM model using binary class dataset.

Graphical representations of the training, validation, and loss accuracies of the CNN-BiLSTM model are presented in Figure 5. On the *y*-axis is displayed the percentage of the dataset that has been given the appropriate labels. The effectiveness of the validation system serves as a benchmark against which the reliability of the training system may be measured. We were able to identify a shift in the process of system optimization that resulted in a spectacular increase in accuracy to 20 epochs. This was achieved as a result of the adjustment. During the validation process, the CNN-BiLSTM model observed an increase in its performance, going from 75 to 84%. A categorical cross-entropy function was utilized so we could assign a numerical value to the training losses that are associated with the suggested system. After 20 epochs, the validation losses were lower, coming in at 0.45 rather than 0.35.

Figure 6 offers an idea of how well the proposed model functions in detecting cyberbullying in the validation process. These studies were conducted with the intention of determining how effectively the proposed paradigm operated in practice. After the beginning, with an accuracy of 94%, the XLM-Roberta model increased its performance over the course of 20 epochs to achieve a validation accuracy of 95%. Measurements of cross-entropy were utilized to reduce the validation loss to a value as low as 0.18, which is a significant reduction from its initial value of 0.14.

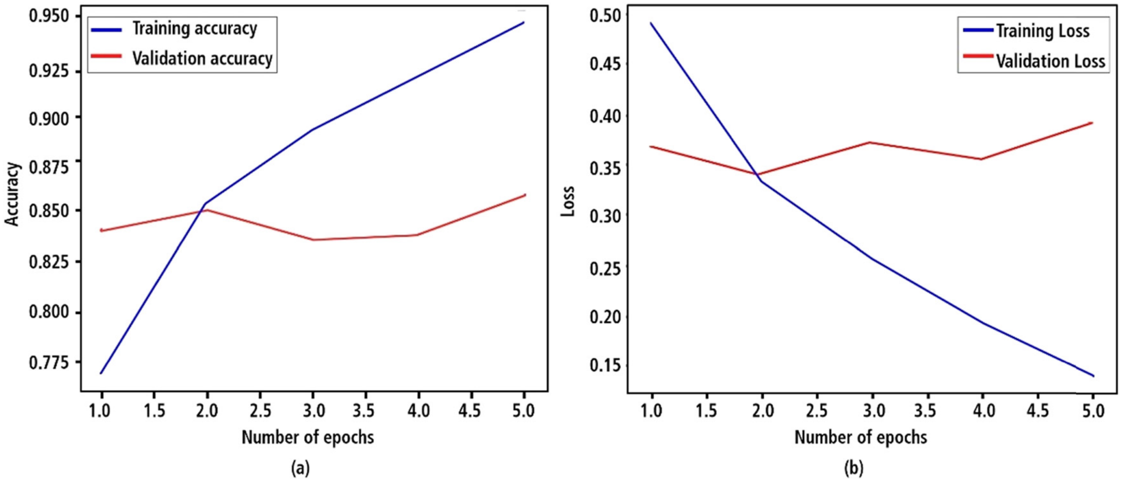

Figure 6: (a) Training accuracy and (b) loss of the XLM-Roberta model using binary class dataset.

4.2 Results of the multiclass classification

In this experiment, the multiclass twitter dataset was evaluated for cyberbullying detection. Based on tweet embeddings, the CNN-BiLSTM model was trained with 70% of tweets, 10% validated with 10% and tested with 20% tweets of the dataset. It classified the dataset into five classes, namely, age, religion, ethnicity, gender, and non-bullying tweets. Tables 3 and 4 summarize the evaluation metric values for the testing results of the CNN-BiLSTM and XLM-Roberta models using multiclass cyberbullying classification.

Figure 7 indicates the confusion matrix for the testing set of the CNN-BiLSTM for the multiclass classification, while Figure 8 shows the confusion matrix of the XLM-Roberta model for the multiclass cyberbullying dataset; out of 7,176 tweets as the testing set, it was observed that 719 were misclassified by the model.

Learning curves are charts that indicate the performance of a model while it is learning over datasets. In this article, we use learning curves as a diagnostic tool to evaluate how well the CNN-BiLSTM and XLM-Roberta models, which learn to train the multiclass dataset, perform throughout the training and validation processes. This model acquires knowledge about the training dataset in small, incremental chunks. As can be seen in Figure 9, the CNN-BiLSTM model improved its training performance from 80 to 94% when using text data and

Table 3: Testing results of the CNN-BiLSTM model based on the multiclass twitter dataset

Label name	Precision (%)	Recall (%)	*F*1-score (%)	Accuracy (%)
Religion	97	95	96	91
Age	98	97	98	
Gender	87	86	87	
Ethnicity	80	80	80	
Non-bullying	92	96	94	

Table 4: Results of the multiclass cyberbullying classification using the XLM-Roberta model

Label name	Precision (%)	Recall (%)	*F*1-score (%)	Accuracy (%)
Religion	83	90	87	94
Age	95	87	91	
Gender	97	96	96	
Ethnicity	100	98	99	
Non-bullying	95	99	97	

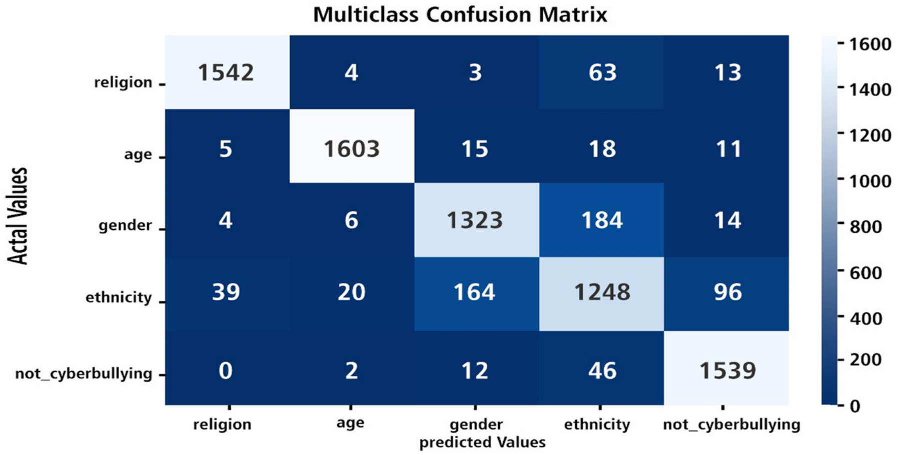

Figure 7: Confusion matrix of the CNN-BiLSTM model for multiclass cyberbullying twitter dataset.

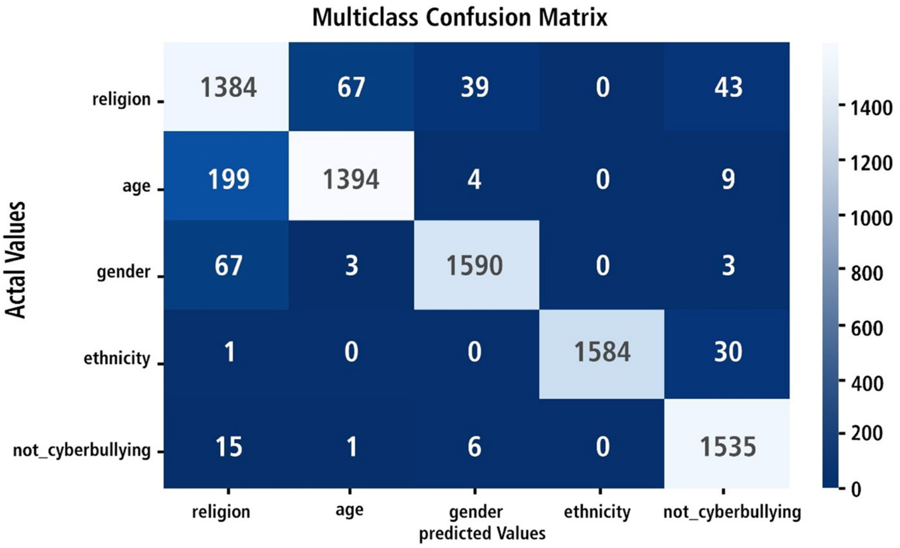

Figure 8: Confusion matrix of the XLM-Roberta model for multiclassification.

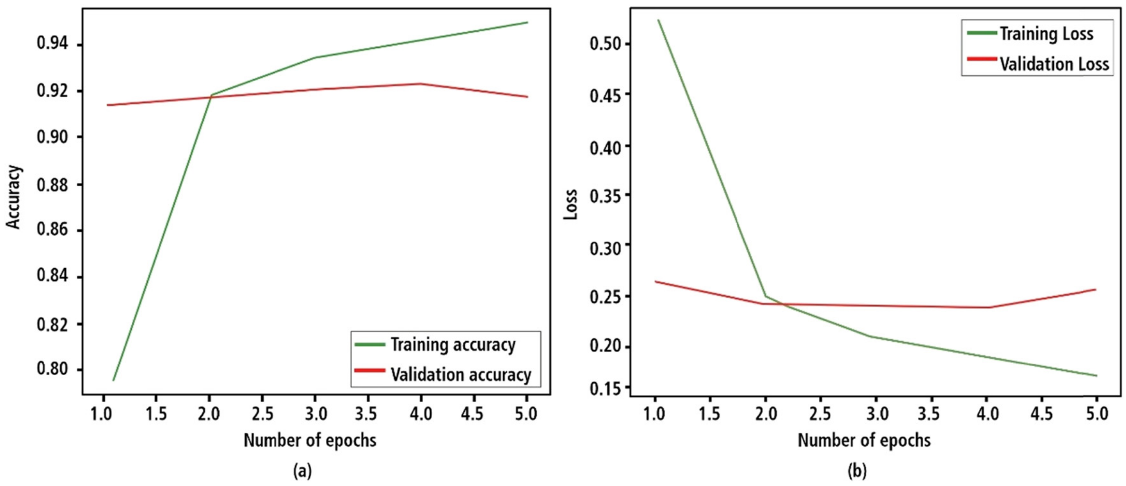

Figure 9: (a) Training and (b) loss performance of the CNN-BiLSTM model using the multiclass class dataset.

its validation performance from 90% when using the same data, with training and validation losses dropping from 0.50 and 0.15 to 0.30 and 0.23, respectively.

The training, validation, and loss accuracy performance of the XLM-Roberta algorithms are shown in Figure 10. After 20 epochs of training, the CNN-BiLSTM model improved its accuracy in the training phase to 99.50%, while its accuracy in the testing phase improved from 92.5 to 94%. The accuracy loss of XLM-Roberta decreased to 0.22.

5 Discussion

Utilizing social media platforms, such as Facebook and Twitter, comes with numerous advantages, but it also involves some disadvantages to consider as well. One of the issues that surfaced as a result of the use of social networks is cyberbullying. It is difficult to put a number on the amount of harm that can be inflicted on

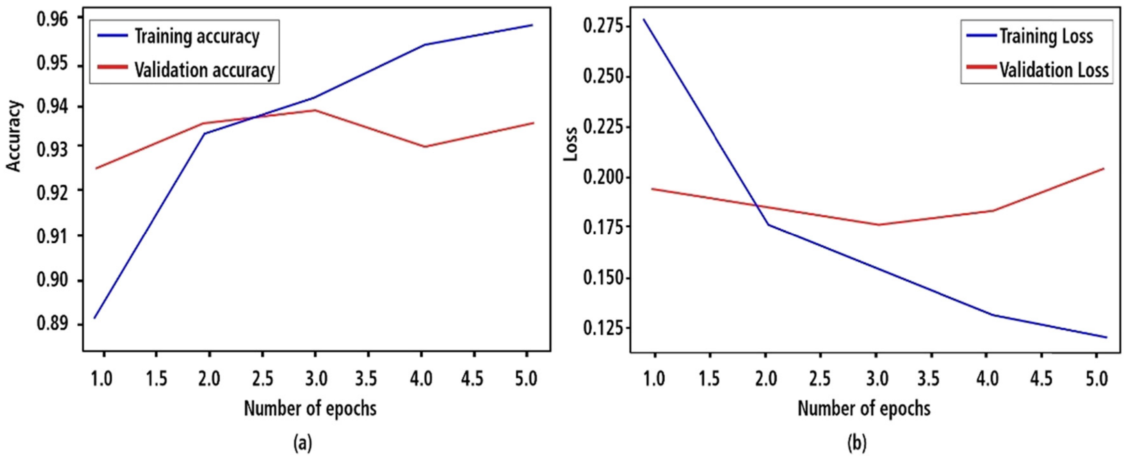

Figure 10: (a) Training and (b) loss performance of the XLM-Roberta model using the multiclass dataset.

Table 5: Results of the proposed system for XLM-Roberta and CNN-BiLSTM compared to existing approaches

References	Preprocessing	Algorithm	Accuracy %
Ref. [37]	Word2vec	Graph convolutional networks	87
Proposed work	Keras embedding	XLM-Roberta	94
		CNN-BiLSTM	91

a victim's life through cyberbullying because of the individual nature of how each person might react to being bullied online. On the other hand, the message may appear relatively normal to those who are not directly affected by it, even though it may be intimidating to the people to whom it is directed. Due to the obscurity of messages that constitute cyberbullying, it is particularly challenging to locate harmful content within them.

Following a discussion and analysis of the testing results from Tables 2–4, we observed that the state-of-the-art XLM-Roberta transformer model proved its effective performance and reported the best results for cyberbullying detection in both used datasets compared to the hybrid deep learning CNN-BiLSTM model, which achieved a lower performance in the case of the Twitter dataset and reasonable results with the Kaggle online discussion dataset. While comparing the results of evaluation metrics for multiclass cyberbullying detection, the XLM-Roberta model provided the highest classification results for ethnicity compared to the other classes, where the CNN-BiLSTM model detected a religious cyberbullying class with higher results than the XLM-Roberta model.

This subsection reports a comparative analysis of the results obtained by the proposed models and study, which are presented in Table 5. Using the accuracy metric and the same dataset, for example, Wang et al. (2020) presented a graph CNN using the Word2vec embedding approach for the detection and identification of various cyberbullying types, and the results were 87% accuracy, where our approach reported 94 and 91% accuracy using the XLM-Roberta and CNN-BiLSTM models correspondingly.

The better performance of the XLM-Roberta model was expected because it transforms and gives the word embeddings of all texts samples presented in the datasets compared to the CNN-BiLSTM model, which consists of an individual embedding neural network layer that was trained with a set of 15,000 word embeddings to create a weight matrix that contains input sequence length and vocabulary size for output embedding a vector to the next layer in the network.

6 Conclusions

This research work presented a CDS that utilizes NLP methods (Word2Vec word embedding) and a supervised hybrid deep learning model (CNN-BiLSTM) to identify online cyberbullying perfectly along with the type of cyberbullying, such as hate speech, gender, ethnicity, religion, and age. To ensure precise cyberbullying detection on social media platforms, this work has used two datasets, Kaggle online discussions and Twitter datasets, to evaluate the proposed system using two scenarios: experiments that were carried out using binary and multiclass cyberbullying detection. While evaluating process using different standard measurement metrics, such as precision, recall, $F1$ score, and accuracy, it was noted that the XLM-Roberta model provided a satisfactory performance, with an 84% accuracy in the binary class Kaggle online discussion dataset and a 94% accuracy using the Twitter multiclass dataset. In future work, the authors will focus on attaining higher accuracy by combining user and network features, as this research is still undergoing in its early stages.

Acknowledgement: This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [NA000230].

Funding information: This research and the APC were funded by the Deanship of Scientific Research at King Faisal University for the financial support under Grant No. NA000230.

Author contributions: Fawaz Waselallah Alsaade and Mohammed Saeed Alzahrani have read and agreed to the published version of the manuscript.

Conflict of interest: Authors state no conflict of interest.

Informed consent: Not applicable.

Data availability statement: The data presented in this study are available here: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset.

References

[1] E. Englander, E. Donnerstein, R. Kowalski, C. A. Lin, and K. Parti, *Defining cyberbullying*, Pediatrics **140** (2017), no. 1, S148–S151.

[2] L. Johnson, *Counselors and cyberbullying: guidelines for prevention, intervention, and counseling*, Retrieved January **7** (2011), no. 2, 2015.

[3] J. Wang, K. Fu, and C. T. Lu, Sosnet: a graph convolutional network approach to fine-grained cyberbullying detection, In *2020 IEEE International Conference on Big Data*, 2020, pp. 1699–1708.

[4] R. Slonje, and P. K. Smith, *Cyberbullying: another main type of bullying*, Scand. J. Psychol. **49** (2020), no. 2, 147–154.

[5] D. Chaffey, *Global social media research summary, smartinsight*, vol. 22, 2020, p. 5.

[6] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, and S. Mishra, Towards understanding cyberbullying behavior in a semi-anonymous social network, *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining Beijing, China*, 2014, pp. 244–252.

[7] S. Cook, *Cyberbullying facts and statistics for 2020, cyberbullying-statistics*, Broadbandsearch, United States, vol. 2, 2020, p. 3.

[8] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, *Detection of harassment on web 2.0.*, In Proceedings of the Content Analysis in the 2020, 2020, pp. 1–7.

[9] K. Reynolds, A. Kontostathis, and L. Edwards, *Using machine learning to detect cyberbullying*, In 2011 10th International Conference on Machine Learning and Applications and Workshops, Pasadena, California, vol. 2021, 2011, pp. 241–244.

[10] A. Kumar, and N. Sachdeva, *Cyberbullying detection on social multimedia using soft computing techniques: A meta-analysis*. Multimed. Tools Appl. **78** (2019), pp. 23973–24010.

[11] K. Dinakar, R. Reichart, and H. Lieberman, *Modeling the detection of textual cyberbullying*, In Proceedings of the social mobile web, Santiago Chile, 2014, pp. 1–10.

[12] M. Dadvar, F. Jong, R. Ordelman, and D. Trieschnigg, *Improved cyberbullying detection using gender information*, In Proceedings of the twelfth Dutch-Belgian information retrieval workshop, University of Ghent, 2012.

[13] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, *Detecting cyberbullying: query terms and techniques*, In Proceedings of the 5th annual, ACM Web Science Conference, Paris, France, 2013, pp. 195–204.

[14] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, *Extracting patterns of harmful expressions for cyberbullying detection*, In Proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'15), The First Workshop on Processing Emotions, Decisions and Opinions, Poznań, Poland, 2015, pp. 370–375.

[15] R. Fiebrink, and M. Gillies, Introduction to the special issue on human-centered machine learning. *ACM T Interactive Intel Syst (TiiS)*. **8** (2018), no. 2, 1–7.

[16] A. M. G. Gualdo, S. C. Hunter, K. Durkin, P. Arnaiz, and J. Maquilón, *The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role*, J. Comput. Educ. **182** (2015), 228–235.

[17] D. L. Hoff, and N. Sidney, *Cyberbullying: causes, effects, and remedies*, J. Educ. Adm. **45** (2009), no. 5, 1–11.

[18] R. Dredge, J. Gleeson, and X. De La, *Presentation on facebook and risk of cyberbullying victimisation. computers in human behavior*, J. Comput. Hum. Behav. **40** (2014), no. 8, 16–22.

[19] S. Ozel, A. Saraç, E. Akdemir, and H. Aksu, *Detection of cyberbullying on social media messages in Turkish*, In International Conference on Computer Science and Engineering (UBMK), IEEE, 2017, pp. 366–370.

[20] W. Romsaiyud, K. Nakornphanom, P. Prasertsilp, P. Nurarak, and P. Konglerd, *Automated cyberbullying detection using clustering appearance patterns*, In 2th International Conference on Knowledge and smart Technology (KST), Riyad, Saud Arabia, 2017, pp. 242–247.

[21] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, and G. Stringhini, *Detecting aggressors and bullies on twitter*, In International World Wide Web Conference 2017, WWW 2017 Companion. International World Wide Web Conferences Steering Committee, New York, USA, vol. 12, 2019, pp. 767–768.

[22] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, *Hierarchical attention networks for cyberbullying detection on the instagram social network*, In Proceedings of the 2019 SIAM International Conference on Data Mining, Calgary, Alberta, Canada, 2019, pp. 235–243.

[23] J. Hani, N. Mohamed, M. Ahmed, Z. Emad, E. Amer, and M. Ammar, *Social media cyberbullying detection using machine learning*, Int. J. Adv. Comput. Sci. Appl. **10** (2019), no. 5, 1–15.

[24] K. Goswami, Y. Park, and C. Song, *Impact of reviewer social interaction on online consumer review fraud detection*, J. Big Data **4** (2017), no. 1, 1–19.

[25] V. Nahar, S. Unankard, X. Li, and C. Pang, *Sentiment analysis for effective detection of cyberbullying*, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, LNCS, vol. 7235, Springer, Berlin, Heidelberg, 2012, pp. 767–774.

[26] Shruthi and C. Mangala, *A framework for automatic detection and prevention of cyberbullying in social media*, Int. J. Innovative Res. Comput. Commun. Eng. **5** (2017), 86–90.

[27] L. Li, B. Qin, B. W. Ren, and T. Liu, *Document representation and feature combination for deceptive spam review detection*, Neurocomputing **254** (2016), 33–41.

[28] T. H. Aldhyani, M. Alrasheed, M. Y. Alzahrani, and H. Ahmed, *Deep learning and Holt-trend algorithms for predicting COVID-19 pandemic*, medRxiv **6** (2020), 1–30.

[29] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, *What yelp fake review filter might be doing*, In Proceedings of the International AAAI Conference on Web and Social Media, Massachusetts, 2013, pp. 409–418.

[30] S. N. Alsubari, S. N. Deshmukh, M. H. Al-Adhaileh, F. W. Alsaade, and T. H. Aldhyani, *Development of integrated neural network model for identification of fake reviews in E-commerce using multidomain datasets*, Appl. Bionics Biomech. **11** (2021), 5522572.

[31] M. E. Alzahrani, T. H. Aldhyani, S. N. Alsubari, M. M. Althobaiti, and A. Fahad, *Developing an intelligent system with deep learning algorithms for sentiment analysis of E-commerce product reviews*, Comput. Intell. Neurosci. **10** (2022), 3840071.

[32] T. H. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, *Cyberbullying identification system based deep learning algorithms*, Electronics **11** (2022), 3273.

[33] H. Alkahtani, and T. H. Aldhyani, *Intrusion detection system to advance internet of things infrastructure-based deep learning algorithm*, Complexity **2021** (2021), 5579851.

[34] M. H. Al-Adhaileh, T. H. H. Aldhyani, and A. D. Alghamdi, *Online troll reviewer detection using deep learning techniques*, Appl. Bionics Biomech. **2022** (2020), 4637594.

[35] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, *Pre-trained models for natural language processing: a survey*, Sci. China Technol. Sci., **63** 2020, 1–26.

[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov, *RoBERTa: A robustly optimized BERT pretraining approach*, arXiv preprint, arXiv: 2019.1907.11692.

[37] J. Wang, K. Fu, and C. T. Lu, *SOSNet: a graph convolutional network approach to fine-grained cyberbullying detection*. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020, pp. 1699–1708.