electronics

MDPI

Article

Cyberbullying Identification System Based Deep Learning Algorithms

Theyazn H. H. Aldhyani [1,*], Mosleh Hmoud Al-Adhaileh [2] and Saleh Nagi Alsubari [3]

[1] Applied College in Abqaiq, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia
[2] Deanship of E-Learning and Distance Education, King Faisal University, P.O. Box 4000, Al-Ahsa 31982, Saudi Arabia
[3] Department of Computer Science, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad 431004, India
* Correspondence: taldhyani@kfu.edu.sa

Abstract: Cyberbullying is characterized by deliberate and sustained peer aggression, as well as a power differential between the victim and the perpetrators or abusers. Cyberbullying can have a variety of consequences for victims, including mental health problems, poor academic performance, a tendency to drop out of work, and even suicidal thoughts. The main objective of this study was to develop a cyberbullying detection system (CDS) to uncover hateful and abusive behaviour on social media platforms. Two experiments were carried out to train and test the proposed system with binary and multiclass cyberbullying classification datasets. Hybrid deep learning architecture consisting of convolutional neural networks integrated with bidirectional long short-term memory networks (CNN-BiLSTM) and single BiLSTM models were compared in terms of their ability to classify social media posts into several bullying types related to gender, religion, ethnicity, age, aggression, and non-cyberbullying. Both classifiers showed promising performance in the binary classification dataset (aggressive or non-aggressive bullying), with a detection accuracy of 94%. For the multiclass dataset, BiLSTM outperformed the combined CNN-BiLSTM classifier, achieving an accuracy of 99%. A comparison of our method to the existing method on the multiclass classification dataset revealed that our method performed better in detecting online bullying.

Keywords: deep learning model; social media; cyberbullying detection system; artificial intelligence algorithms

Citation: Aldhyani, T.H.H.; Al-Adhaileh, M.H.; Alsubari, S.N. Cyberbullying Identification System Based Deep Learning Algorithms. *Electronics* **2022**, *11*, 3273. https://doi.org/10.3390/electronics11203273

Academic Editors: Domenico Ursino and Arkaitz Zubiaga

Received: 19 August 2022
Accepted: 10 October 2022
Published: 12 October 2022

1. Introduction

Most people now use social media to communicate on a daily basis, and the use of social media has spread across all ethnicities and demographic groups. Due to the pervasiveness of social media and the relative anonymity it provides, cyberbullying has the potential to harm anybody anywhere and at any time. On 15 April 2020, UNICEF posted a notice regarding the heightened risk of online harassment and bullying during the COVID-19 outbreak as a result of school suspensions, increased screen usage, and decreased face-to-face interactions. Cyberharassment is defined as the use of electronic communication to intimidate a person or group of persons online, typically by sending messages of an intimidating or threatening nature [1]. Cyberbullying is associated with traditional bullying and is an important area of research. The growth of online bullying is shocking: in middle school and high school, 36.5% of students say they have been harassed and threatened online, while 87% say they have witnessed harassment. The consequences of cyberbullying for victims can range from poor academic achievement to unhappiness to suicide attempts. Teaching students "internet street smarts", watching for warning indicators, and counselling are the primary approaches used to prevent online harassment [2]. Although cyberbullying is illegal in all 50 states, the bulk of regulations have no authority outside of school. While social media networks such as Facebook, Twitter,

Instagram, Snapchat, and others offer cyberbullying instructions and materials, they do not offer active anti-cyberbullying tools. Approximately 90% of cyberbullying incidents go unreported. Thus, an intelligent, functioning, user-generated text detection system is essential. Even though numerous groups are dedicated to increasing awareness about cyberbullying, the number of digital attacks continues to rise [3]. Three billion people utilize social network platforms to communicate with others [4]. Facebook and other social networking apps are clearly beneficial to their users, but they can also be used for harmful purposes. The employment of digital technology to abuse a person or group of people is considered online harassment, and can be considered a malign use of technology [5]. Since cyberbullying can swiftly reach a large number of individuals, it has more potent and long-lasting impacts than traditional bullying. Moreover, it can be difficult or even impossible to remove hazardous information from online sources. Although there is no evidence that online bullying causes physical injury to victims, mental health issues, such as despair, low self-esteem, fatigue, and even suicide attempts, have been linked to cyberbullying [6]. Especially among youngsters and teenagers, cyberbullying has grown in prevalence over the last decade. A recent research study indicated that 37% of children in India experienced cyberbullying in 2018, followed by the United States (26%), South Africa (26%), and Turkey (20%). Research findings indicate that this problem is increasing rapidly and is unrelated to a country's level of growth. Between 2011 and 2018, cyberbullying increased considerably in Sweden, one of the world's most developed countries [7]. A considerable amount of research has applied machine learning techniques to automatically identify cyberbullying [8,9], although the vast majority of studies have been conducted in English. Text mining techniques, such as those utilized in sentiment analysis studies, have been used in most of the research in this area. Posts on social media, by their very nature, are dynamic and context-dependent, and thus they should not be considered as standalone texts [10].

In this research, we have aimed to develop a hybrid deep learning model based on a convolutional neural network and bidirectional long short-term memory (which are artificial neural network techniques) to detect and predict different types of content containing cyberbullying activities and behaviors on social media platforms (e.g., Twitter and online discussion blogs). These different types of content are associated with religion, age, gender, ethnicity, and aggressive or hateful speech.

2. Literature Review

Yin et al. [8] carried out the first research into the automatic recognition of online cyberbullying. The authors used three different datasets to detect harassment on three different online platforms. The Kongregate platform was used for one dataset collection, while the other datasets were gathered from discussion-based communities (e.g., Reddit). A linear kernel classification model and various feature extraction methods (N-grams and term frequency-inverse term frequency (TF-IDF)) were employed for the classification task. Although their experimental results were ambiguous, the study served as a starting point for further investigation. Another study was proposed in the same field by [9]. The authors implemented C4.5, k-nearest neighbors (KNN) and support vector machine (SVM) classification techniques, which were tested on a dataset consisting of text comments collected from the Formspring.me platform. Based on their experimental results, the C4.5 decision tree algorithm surpassed both the KNN and SVM classifiers, with a detection rate accuracy of 78.5%. Dinakar et al. [11] proposed a two-step detection method. The very first step was to decide whether or not a piece of information or content falls under the category of sensitive. The second step involved classifying the content of the text with a particular label (e.g., intellectual ability or sexual orientation). The proposed method was tested on 4500 YouTube comments, and the classification accuracy ranged from 70 to 80%. Dadvar et al. [12] proposed a gender-based method to detect cyberbullying related to gender harassment. Their approach employed two distinct vocabulary sets. Based on their findings, this method has marginally enhanced the accuracy of machine-learning

classifiers. Subsequently, several studies using a range of different techniques have been conducted in relation to cyberbullying detection. Based on Essential Dimensions of Latent Semantic Indexing (EDLSI), Kontostathis et al. [13] developed a model for classifying the most popular words used in cyberbullying based on messages from the Formspring.me website. The authors reported that the classification model provided an average precision of 91.25%. Ptaszynski [14] used brute force search algorithms and learning classifiers to find patterns associated with online cyberbullying. Specifically, in their classification process they extracted patterns from sentences. Based on the Human Rights Center database, this approach surpassed earlier cyberbullying detection methodologies. Zhang et al. [15] used deep learning to design a robust cyberbullying identification model. A convolutional neural network (CNN) model was built using the pronunciation of words as input features for the detection process. The CNN model was tested and verified on a dataset consisting of social media text comments gathered from the Twitter and Formspring.me platforms. The results showed that the pronunciation-based CNN model performed better than baseline CNN models with arbitrarily created word embeddings. Chavan and Shyla [16] presented a method for determining whether a comment would be insulting to other users. They used skip-grams as input sequences for their machine learning classifiers. Furthermore, they incorporated the results of SVM and logistic regression classification models into their methodology. Squicciarini et al. [17] used a decision tree classifier to identify text-based features and then presented a rule-based method to further identify cyberbullying behaviors.

According to the literature review on cyberbullying detection, few studies have focused on analyzing texts written in languages other than English. Among the studies that have, Ozel et al. [18] used the Turkish language in their investigation of cyberbullying detection. To generate an evaluation dataset for their experimental work, they collected streaming data from Twitter. Each tweet was given its own vector using the bag-of-words approach and classified using a variety of machine learning techniques (support vector machine, naïve Bayes, C4.5 and KNN) to determine whether the posts involved mistreatment. In terms of F-measure, the Naive Bayes classifier significantly outperformed other classification techniques, with an accuracy rate of 79%.

Bozyigit et al. [19] used data from to create a Turkish dataset for identifying cyberbullying. The authors applied different neural network techniques in their experiments. To reduce feature space dimensions and eliminate unnecessary words, the information gained was used to determine the importance and ranking of features before the models were trained. The F-measure was 91% for this artificial neural network technique. Wang et al. [3] presented a graph convolutional neural network technique (GCN) for multi-class cyberbullying detection using 40,000 Twitter posts. Further, the authors compared various machine learning techniques, such as XGBoost, Naïve bays (NB), SVM, MLP, and KNN. Based on their experimental results, the GCN model achieved the highest F1-score of 92% [20]. Another study presented by Bozyiğit et al. [20] used supervised machine learning techniques (SVM, logistic regression, NB, random forest and AdaBoost) to detect cyberbullying based on numerical and text-based features. Out of the algorithms used, the AdaBoost algorithm had the best performance.

Recently, there has been a great deal of interest expressed by the academic community in the topic of cyberbullying. In this section, we discuss the contributions made by academics to research on the detection of cyberbullying [21–25]. Hosseinmardi et al. [26] created a methodology for identifying instances of cyberbullying on Instagram by mining the platform's captions, comments, and photo data. The information was generated with the use of the Instagram's API, as well as user profiles. In addition, the dataset obtained was annotated with the help of the CrowdFlower platform. Logistic regression was used to predict bullying postings on the set40+ dataset. The criteria that were taken into account were early comments, captions, post times, user traits and photo content. The F1-score of 0.84 was determined by combining the usage of unigram features with those of bigram features. Using a DL model, AlAjlan et al. [27] were able to locate cases of cyberbullying.

Using approaches for feature selection and feature engineering, they were able to extract features from the input data. They removed duplicates from a dataset that had 39,000 tweets before using them in their study. The model was trained and validated using a total of 9000 instances of bullying tweets and 21,000 instances of tweets that did not include bullying. With a rate of 95%, their model was substantially more accurate than the SVM.

We utilized the research carried out in to apply it in a practical setting to Turkish texts, since the detection of cyberbullying has been neglected. To detect instances of cyberbullying in Turkish social media platforms, the authors created eight unique models of artificial neural networks. According to the findings, their machine learning classifiers performed far better than those used in prior tests, achieving a score of 91% on the F1-measure [18]. Aldhyani et al. [28] proposed deep learning models to detect Suicidal Ideation on Social Media. The work carried out in [29] illustrates a situation that is similar to this one. In their investigation, they presented a strategy for recognizing and responding to instances of cyberbullying, with a specific focus on the treatment of content t written in Arabic [29].

3. Materials and Methods

This section describes the key points of the proposed cyberbullying detection system (CDS) framework used for investigating and recognizing cyberbullying activities on different social media platforms (e.g., Wikipedia Talk pages and Twitter). Figure 1 presents the steps that are applied in this framework.

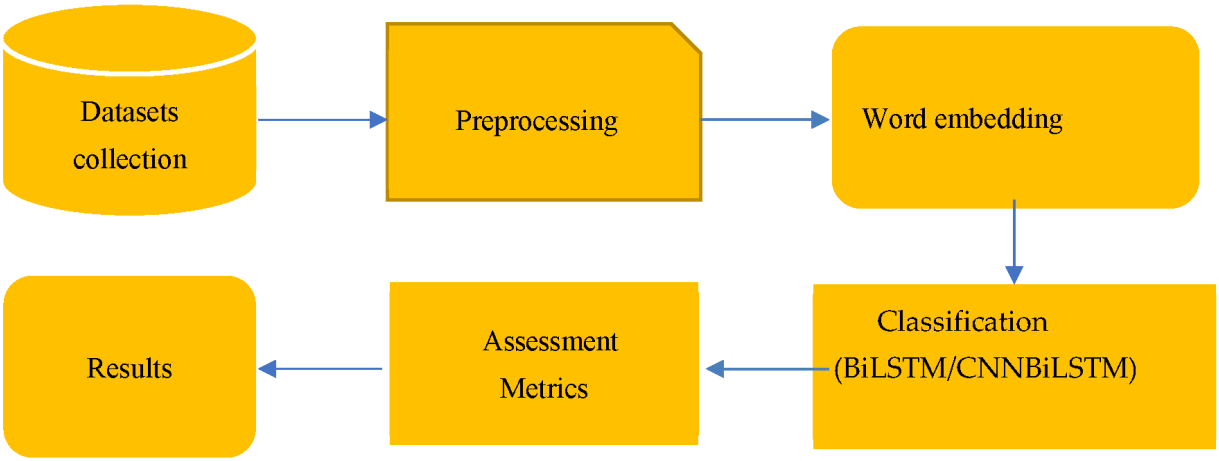

Figure 1. The framework of the applied cyberbullying detection system (CDS).

The details of this framework are discussed below.

3.1. Dataset Collection

This is the most important phase in the proposed methodology. To perform our experimental work on online cyberbullying analysis and detection, we employed two different social media datasets, both of which were collected from the Kaggle platform.

3.1.1. Binary Aggressive Cyberbullying Dataset

As the name suggests, aggressive cyberbullying is a type of abuse or bullying carried out over the Internet. Online bullying is also another term for cyber-harassment. This is a binary dataset consisting of 115,661 post samples, distributed as 101,082 aggressive posts and 14,782 non-aggressive posts, collected from the Wikipedia Talk website [30].

3.1.2. Multiclass Cyberbullying Dataset

This is an openly accessible dataset gathered from the social networking site Twitter, where users share and converse with texts called tweets. The dataset includes 39,869 tweet samples, distributed across five types of online bullying classes, such as religion, age,

gender, ethnicity, and non-cyberbullying tweets [3]. Figure 2 visualizes the dataset sample distribution per class.

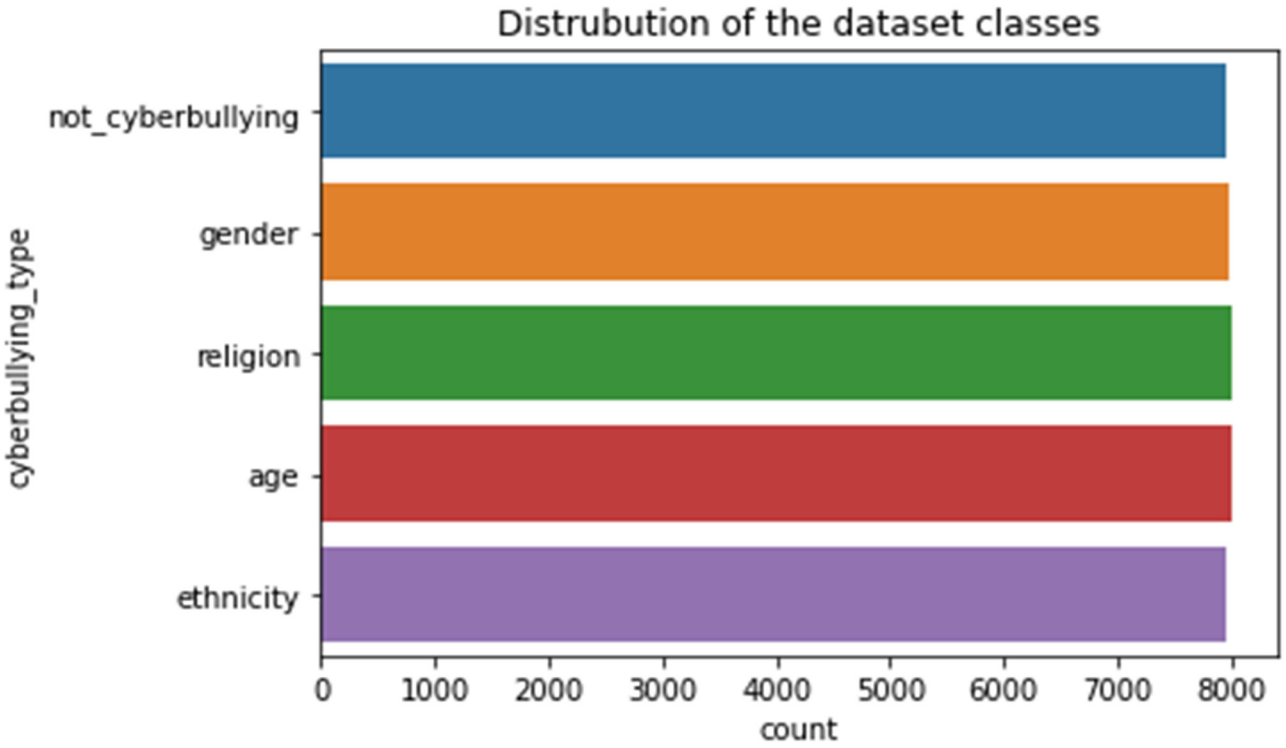

Figure 2. Visualization of the sample distribution for the multiclass dataset.

As illustrated in the above Figure 2, the distribution of dataset tweets in each class can be seen: religion has 7998 samples, age class has 7992 samples, gender has 7973 samples, ethnicity has 7961 samples, and the not_cyberbullying class has 7945 normal tweets.

3.2. Preprocessing

The preprocessing step involves cleaning and removing noise from the dataset before the representation and transformation methods are applied. The goal of text processing is to transform and express online social post content in a way that can be analyzed and classified using the employed deep learning methods. However, the datasets need to be cleaned using the following steps:

- Punctuation removal, which is the act of taking out all of the punctuation (e.g., ?, ! : ; ‘) from each social post in order to make it look cleaner. “The,” “a,” “an,” and “in” are some of the stop words that are removed from a dataset for this purpose.
- Transforming all words in the text into lower case words.
- Removing all unnecessary words, emojis, white spaces, and digits characters from the social posts in the datasets.
- Tokenization, which is the process of breaking a sentence down into its constituent parts, such as words, phrases, and other pieces of information.
- As we used deep learning neural networks techniques to identify each social post as cyberbullying or not, all text sequences in the datasets must have equal real-value vectors. The post padding sequence method is used to complete this task.

3.3. Word Embedding Representation Approache

Word embedding is a technique adopted in various text mining tasks to form vector representations of words of the given text content. When performing text classification, it generally obtains form of a real-valued word vector that embeds the semantic and context

of the words that are relatively close in the vector space and predicted them with similar meanings. For example, word2Vec [31,32] has two types of algorithms (namely skip-gram and 'Continuous Bag of Words') which are Google-developed by using two-layers neural network to predict the context of the given word in the text. While these embedding algorithms cannot handle words that are not in the vocabulary size, it is a common choice for natural language processing tasks [33,34]. In this work, a Keras embedding layer [35] was used on selected vocabularies from the binary and multiclass datasets. We have selected Keras embedding rather than pre-trained embedding models since the first takes less time and easy resource computing in addition it takes all texts (cleaned posts and tweets in this case) and form their word vectors as input data for the proposed models. More details about this layer can be found in the next section of this study.

3.4. Classification Techniques

After performing the preprocessing steps on both cyberbullying datasets, the next step is the classification of the social media posts into various types of online bullying categories. For the detection and classification tasks, we experimented with two dissimilar supervised neural network techniques: CNN combined with bidirectional long short-term memory (CNN-BiLSTM) and single BiLSTM independently.

3.4.1. Bidirectional Long Short-Term Memory

The LSTM network is a type of artificial recurrent neural network that is utilized in a variety of artificial intelligence and deep learning activities, including natural language processing, image processing, sequence mining, and text mining [36,37]. The memory cells employed in the LSTM can transfer the results of prior data features into the output. Furthermore, feature learning occurs in just one direction: forward. This ignores backward construction and hence decreases the performance of the machine learning system. To address this problem, the BiLSTM connects two hidden layers with opposing orientations to a single output, and the data features are then processed and achieved in two directions: forward and backward. The production layer in a network can gain sequential knowledge from history and upcoming states instantly. Figure 3 shows the structure of the BiLSTM model for cyberbullying post content detection using text-based features.

Each LSTM memory unit has four gates: input i_t, forget f_t, cell state c_t, and output gate o_t. The equations of these gates are presented as follows [30].

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

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right) \quad (2)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o \quad (3)$$

$$c_t = f_t c_{t-1} + i_t * tanh(\ W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (4)$$

$$\overrightarrow{h_t} = o_t * tanh(c_t) \quad (5)$$

$$\overleftarrow{h_t} = o_t * tanh(c_t) \quad (6)$$

$$tanh(x) = \frac{e - e^{-x}}{e\ + e^{-x}} \quad (7)$$

$$H_t = \left(\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right) \quad (8)$$

where *sig* and *tanh* represent the sigmoid and tangent activation functions individually, x is the input sequence, W and b specify weight and bias factors, C_t is the cell state, h_t denotes the output of the LSTM cell, and H_t is the output of the bidirectional concatenation of $\overrightarrow{h_t}$ forward and $\overleftarrow{h_t}$ backward LSTM layers at the current time t.

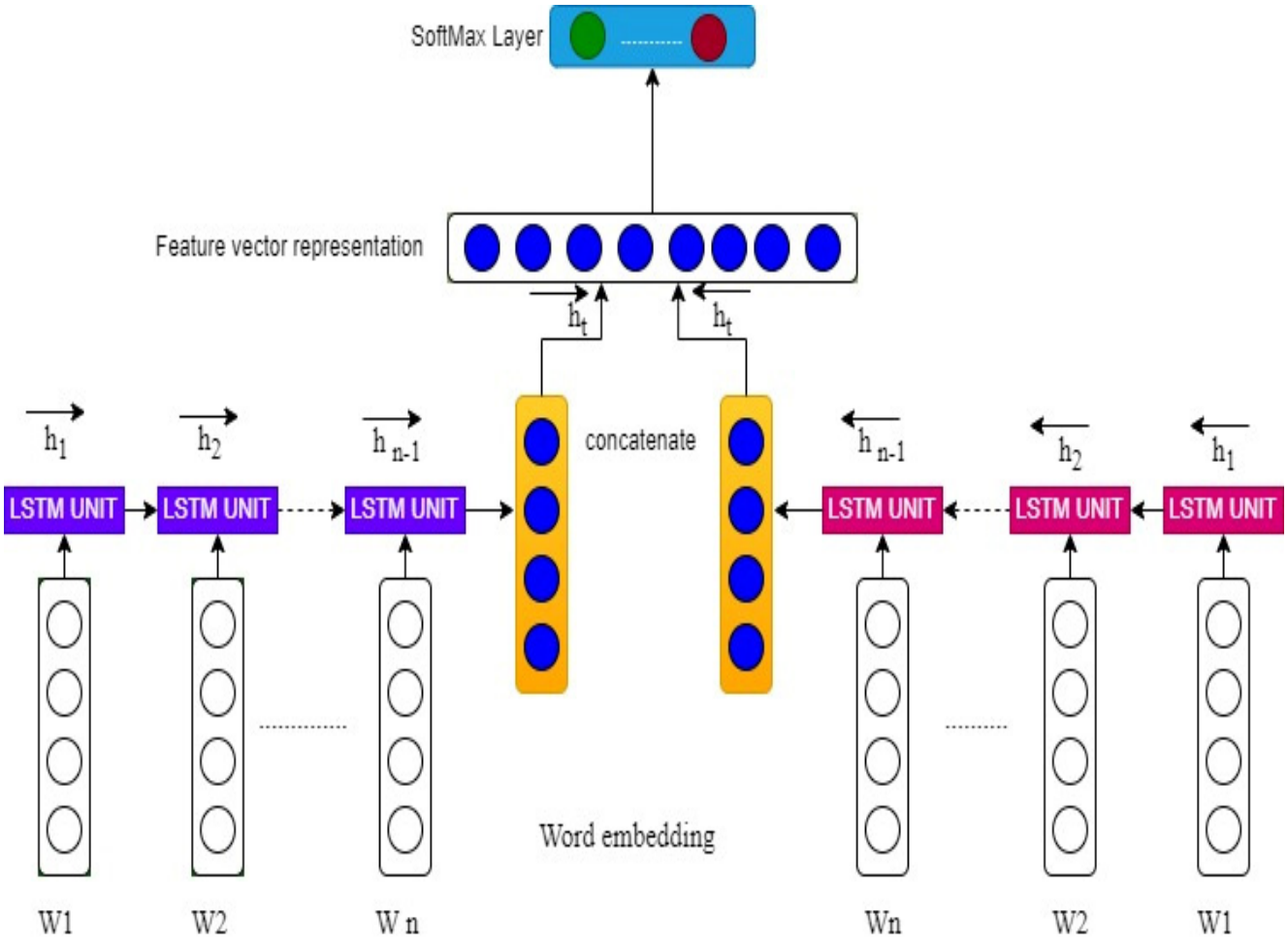

Figure 3. Structure of the BiLSTM model for cyberbullying detection.

As illustrated in Figure 2, the BiLSTM model consists of three hidden neural layers to appropriately interpret the properties of the actual meaning of words, sequences, and phrases in the social media post. The first layer is an embedding layer that is the input layer designed with three components: maximum features, embedding dimension and input sequence length. Maximum features (vocabulary size) are the top 20,000 and 30,000 most frequent words selected from the training dataset of the binary and multiclass datasets respectively, as represented in Figure 3 by W_1, W_2 ... W_n. The embedding dimension determines the size of word embedding vector for each word in the selected vocabulary words that were encoded into sequences of integers. Custom embedding with specified 50 dimensions vector space was applied for word embeddings for the binary and multiclass datasets respectively. The input sequence length is defined as the average length of every input social media post in the datasets and set to 249 and 98 words. The main function of an embedding layer in the BiLSTM model is to make an input embedding matrix for each randomly chosen word from the training set, send it to the forward and backward LSTM (100 units) layers for analysis, and find the semantics of the input sequences of the social post contents so that an output SoftMax layer can classify them into different cyberbullying categories.

$$E(w) = R^{V \times D} \tag{9}$$

where $E(w)$ is the embedding matrix, R is the real number system, V represents the vocabulary size (maximum features), and D refers to the dimension of the word embeddings vector.

3.4.2. Combined Convolutional Neural Network with BiLSTM Model

A CNN is a type of computational intelligence neural network. It is commonly utilized to discover complex patterns in image processing, computer vision, and natural language processing applications [38–40]. A CNN is based on the architecture of the visual cortex and closely matches the human brain's connected configuration of neurons. Convolution in the CNN technique is a crucial component of artificial neural networks, which defined as a mathematical operation passes on an input data matrix of the network There are three main components of the CNN: a convolutional layer, max pooling, and the full-connected layer. In this study, we applied the CNN with BiLSTM to construct an online cyberbullying detection system. Figure 4 depicts the overall structure of the CNN-BiLSTM model.

The following section discusses the components of the CNN-BiLSTM model.

- Embedding layer

This is the first layer of the CNN-BiLSTM model by which an embedding matrix is constructed for each social medial post of the dataset. In particular, it is known as lookup table used to generate and map word embedding of each word in the selected vocabulary from the binary and multiclass datasets into numerical representation form. It consists of three parameters (input sequence length, embedding dimension and vocabulary size), as explained above.

- Convolutional layer

The convolution layer is the most important layer in the CNN structure. It performs mathematical computations on the input embedding matrix provided by an embedding layer. It uses filters to pass across the input word embedding matrix in order to gather sequence information and reduce the dimensions of the input sequence. In this layer, the convolution process is conducted in one dimension. In the CNN-BiLSTM model, we used 100 filters with a window size of three kernels to pass on word-based representations of the social media post contents in order to extract a set of local features of word sequences from an input embedding matrix provided by embedding layer. An Equation (11) is presented for the convolutional operation:

$$F = CV(W, X) \tag{10}$$

where CV represents the number of the convolves, W is the filter map, F is the output feature map of the convolutional process. Then, the ReLU activation function is used for avoiding the overfitting problem in the attained data as expressed as follow.

$$\sigma = Relu(F + b) \tag{11}$$

where b is bias factor. Further, the output of activation function σ are passed into the max pooling layer using the convolution kernel, which extracts important features to improve the classification accuracy. The equation for max pooling layer is as follows:

$$Q_i = \text{Max}\left(P_j^1,\ P_j^2, P_j^3, \ldots., P_j^t\right) \tag{12}$$

where Q_i is the feature vector extracted from the max pool, and P_j^t represents the feature map before the maximization process.

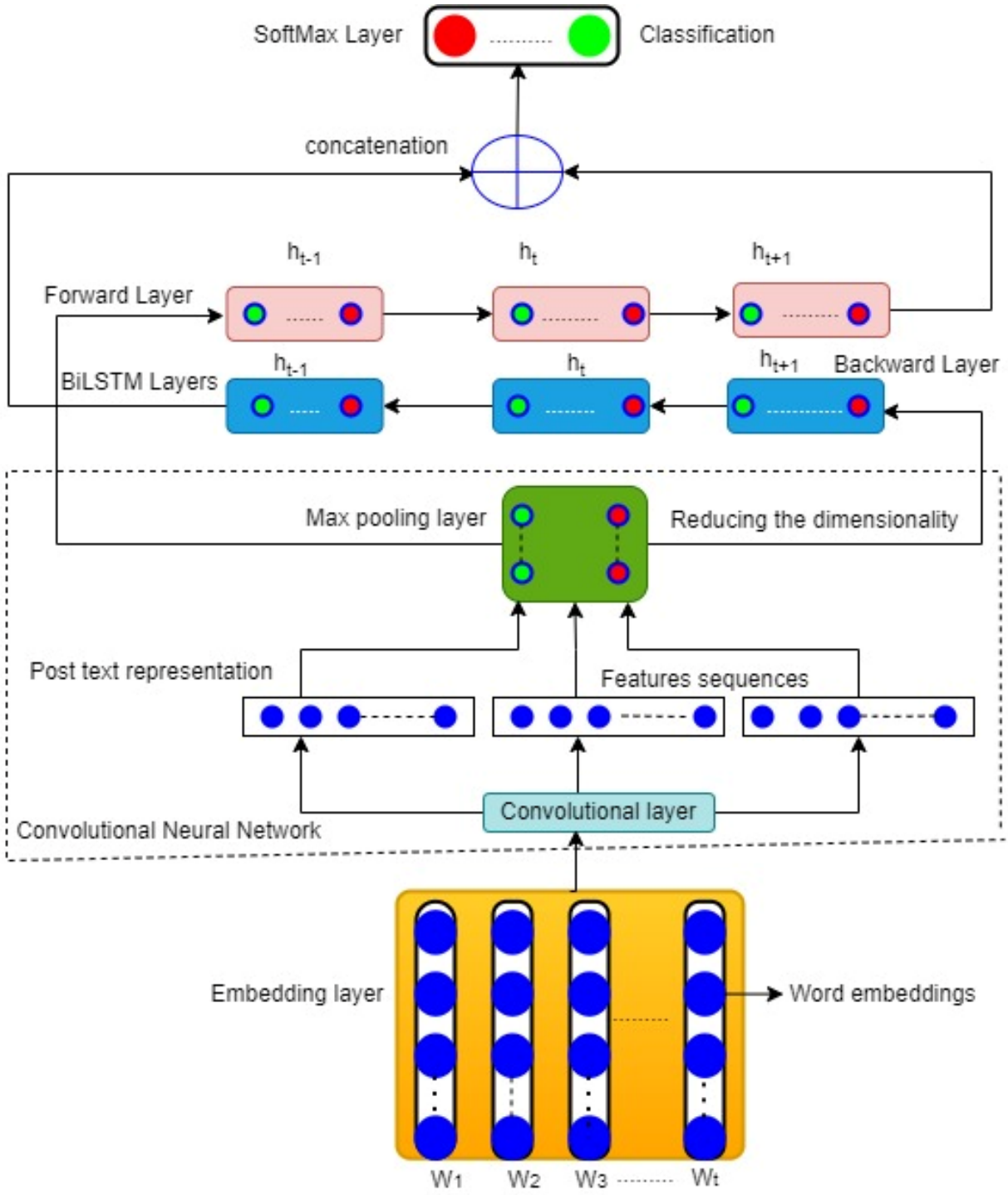

Figure 4. Structure of the CNN-BiLSTM model for cyberbullying detection.

- SoftMax layer

This is the last layer applied in the CNN-BiLSTM model, which is used for the classification of output classes of the evaluated datasets. The number of neurons in this layer is set based on the number of classes in the dataset. We performed two experiments using different cyberbullying detection datasets (binary and multiclass datasets). For this purpose, we placed two and five neurons in this layer independently [41,42]. Furthermore, its activation function performs probability distribution calculations for the input sequence vector of each different type of cyberbullying activity and behavior presented in the dataset, such as

age, gender, aggression, ethnicity, and religion. The equation for the SoftMax activation function is expressed as follows:

$$\sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{13}$$

where z denotes the values of the neurons placed in the output layer, and e is an exponential that acts as non-linear function. The significant parameters of deep learning is shown in Table 1.

Table 1. Summary of the parameters used in the CNN-BiLSTM model.

Parameter Name	Value
Input sequence length	249, 98
Embedding dimension	50
Vocabulary size	20,000, 30,000
Number of filters	100
LSTM units	100
Dropout	0.3
Batch size	128, 10
Number of epochs	5
Activation function	ReLU
Optimizers	RMSprop, Nadam

3.5. Assessment Metrics

This section presents the assessment metrics used to verify and measure the performance of the applied deep learning models: combined CNN-BiLSTM and single BiLSTM for classifying social post content as cyberbullying or non-cyberbullying. We employed various standard evaluation metrics to appraise the suggested models based on a number of false-positive and false-negative samples attained from the confusion matrix illustrated in the next section. These assessments metrics are specificity, recall, precision, F1-score, and accuracy. The following equations are defined for these metrics:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \times 100 \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \times 100 \tag{16}$$

$$specificity = \frac{TN}{TN + FP} \times 100 \tag{17}$$

$$F1 - score = 2 \times \frac{precision \times Sensitivity}{precision + Sensitivity} \times 100 \tag{18}$$

4. Results

This section reports the results of our experiments that were carried out using the deep learning models employed to develop the CDS system for detecting and categorizing associated linguistic cyberbullying into multiple classes. The aggression, age, religion, ethnicity, gender, and non-bullying contents are included in this section. The proposed CDS system was tested in two different settings, binary and multiclass classification, with two separate real cyberbullying datasets. Before performing the classification task, we have

used five-folds cross validation to divide the dataset samples into training, testing, and validation sets, as shown in Table 2.

Table 2. Splitting of the datasets.

Dataset Name	Total of Samples	Training Set 70%	Validation 10%	Testing 20%
Binary class dataset	115,864	83,422	11,586	20,855
Multiclass dataset	39,869	28,705	3987	7176

4.1. Binary Classification Results

The proposed deep learning algorithms were applied to investigate the effectiveness of binary bullying classification. We categorized the binary class dataset into aggressive and non-aggressive bullying classes. The CNN-BiLSTM as well as single BiLSTM were considered as classification models for this dataset. Figure 5 depicts the confusion matrices for the binary classification dataset.

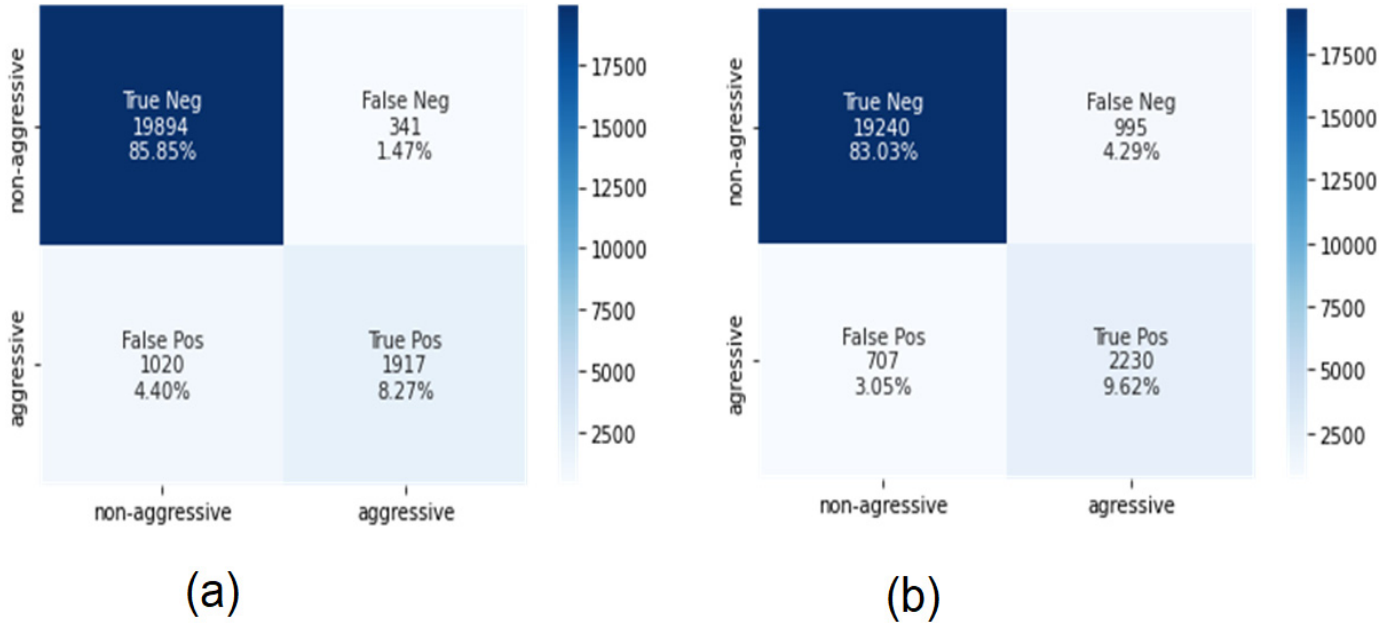

Figure 5. Confusion matrices of the BiLSTM (**a**) and CNN-BiLSTM (**b**) using the binary dataset.

Comparing the misclassification rates, the BiLSTM showed promising results with a 4.40% false-positive rate and a 1.47% false-negative rate, while the CNN-BiLSTM model had a 3% false-positive rate and a 4.29% false-negative rate. Thus, the BiLSTM model achieved better classification results than the CNN-BiLSTM model for cyberbullying detection. Table 3 presents the classification results.

Table 3. Binary classification results of the proposed CNN-BiLSTM and BiLSTM models.

Algorithms	Precision (%)	Recall (%)	Specificity (%)	F-Score (%)	Accuracy (%)
CNN-BiLSTM	69.1	76	95	72.3	93
BiLSTM	85	65.2	98.3	74	94.1

As this dataset has an imbalanced class problem, the F1-score metric is an appropriate evaluation metric for measuring the proposed algorithms. The empirical results revealed that BiLSTM outperformed the CNN-BiLSTM, improving the detection rate by 1% in terms

of the F1-score. Figure 6 demonstrates the receiver operator characteristic (ROC) and precision-recall curves for the binary class dataset using the BiLSTM.

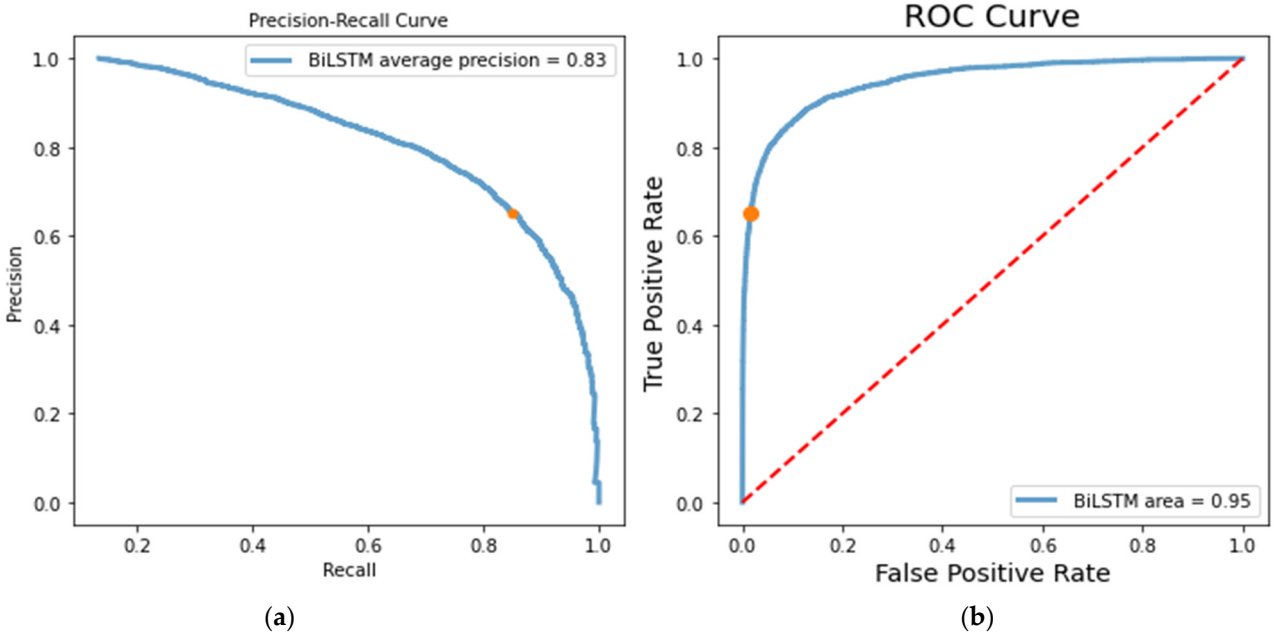

Figure 6. (**a**) precision-recall and (**b**) ROC curve for the binary class dataset using the BiLSTM model.

Figure 7 shows the performance of the CNN-BiLSTM algorithms with respect to training, validation, and loss accuracies using five epochs. The CNN-BiLSTM model initially achieved training accuracy of 96% in the training phase and the model reached testing accuracy value of 92.69% in testing phase after five epochs. The BiLSTM model started with training accuracy of 91.15% and reached to 94.45% in the training phase.

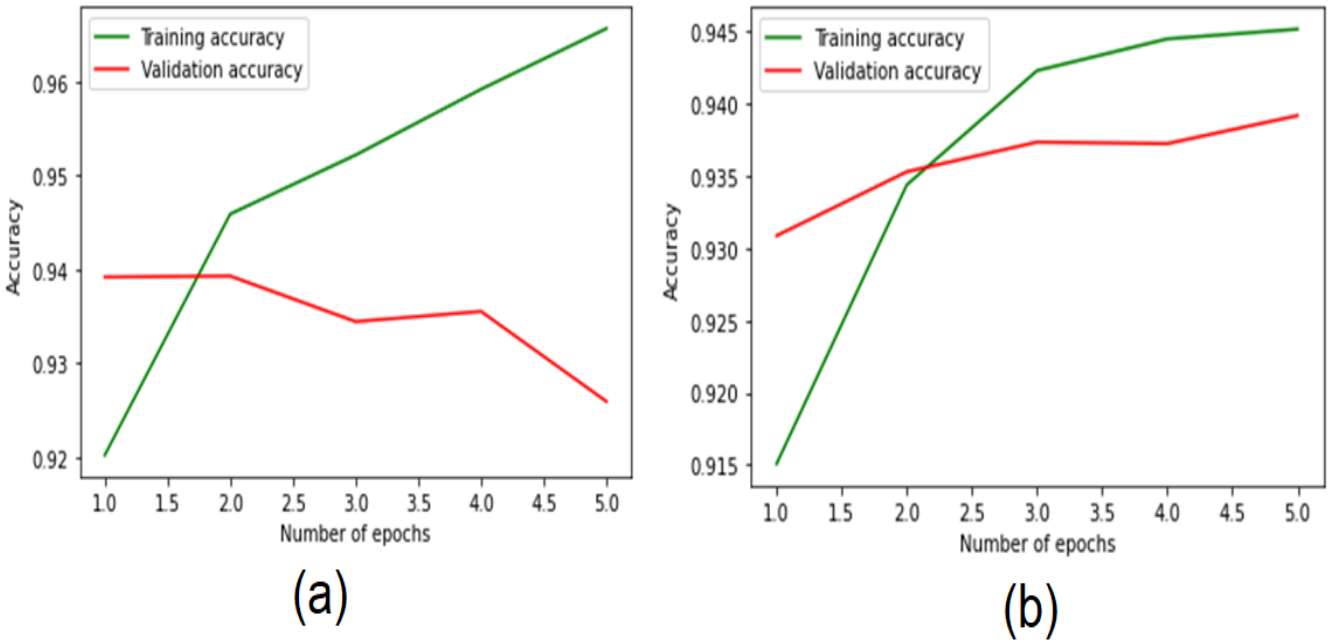

Figure 7. Performance plots of (**a**) CNN-BiLSTM and (**b**) BiLSTM using the binary class dataset.

The accuracy losses of the two models are presented in Figure 8. Both models had difference losses scores, and the BiLSTM had lower loss compared to the CNN-BiLSTM.

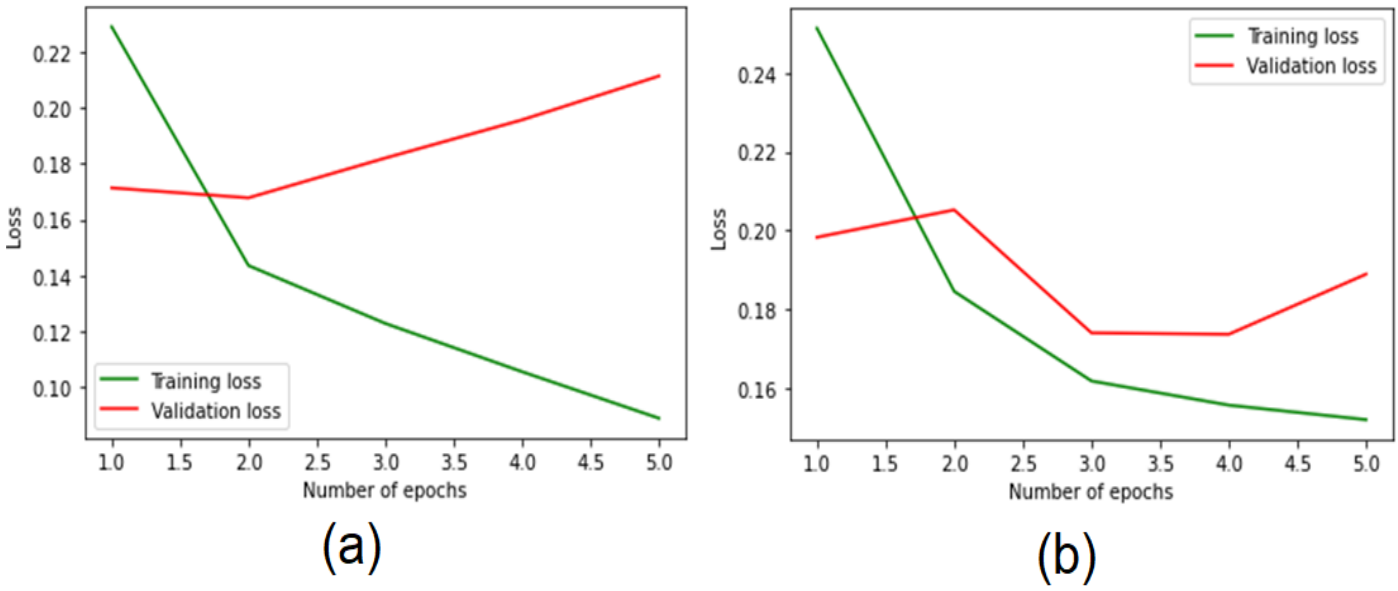

Figure 8. Accuracy loss of the (**a**) CNN-BiLSTM and (**b**) BiLSTM models using the binary dataset.

4.2. Multiclass Classification Results

Based on the learning of dissimilar tweets embeddings, the same proposed deep learning models described above were investigated to find and classify different types of cyberbullying classes (e.g., gender, ethnicity, age, religion, and non-bullying) in the evaluated multiclass dataset. To test the models, we adopted various supervised classification and evaluation metrics, including precision, recall, F1-score, and accuracy. These metrics were calculated using confusion matrices, which are shown in Figures 9 and 10.

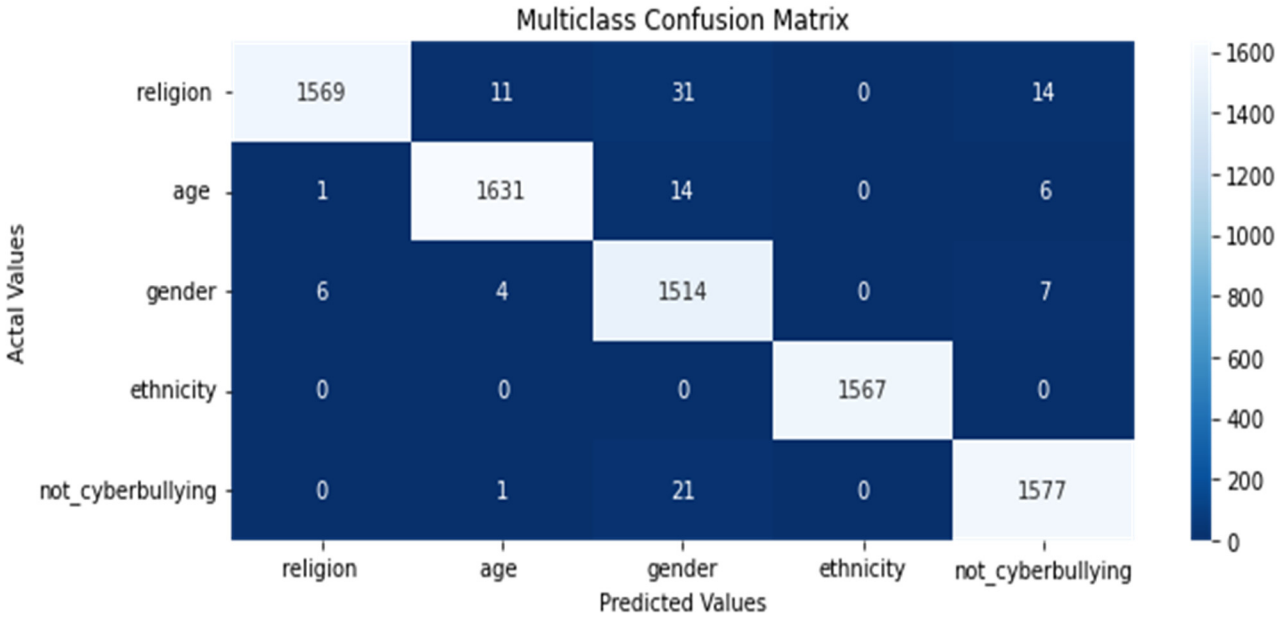

Figure 9. Confusion matrix of the BiLSTM model using the multiclass classification dataset.

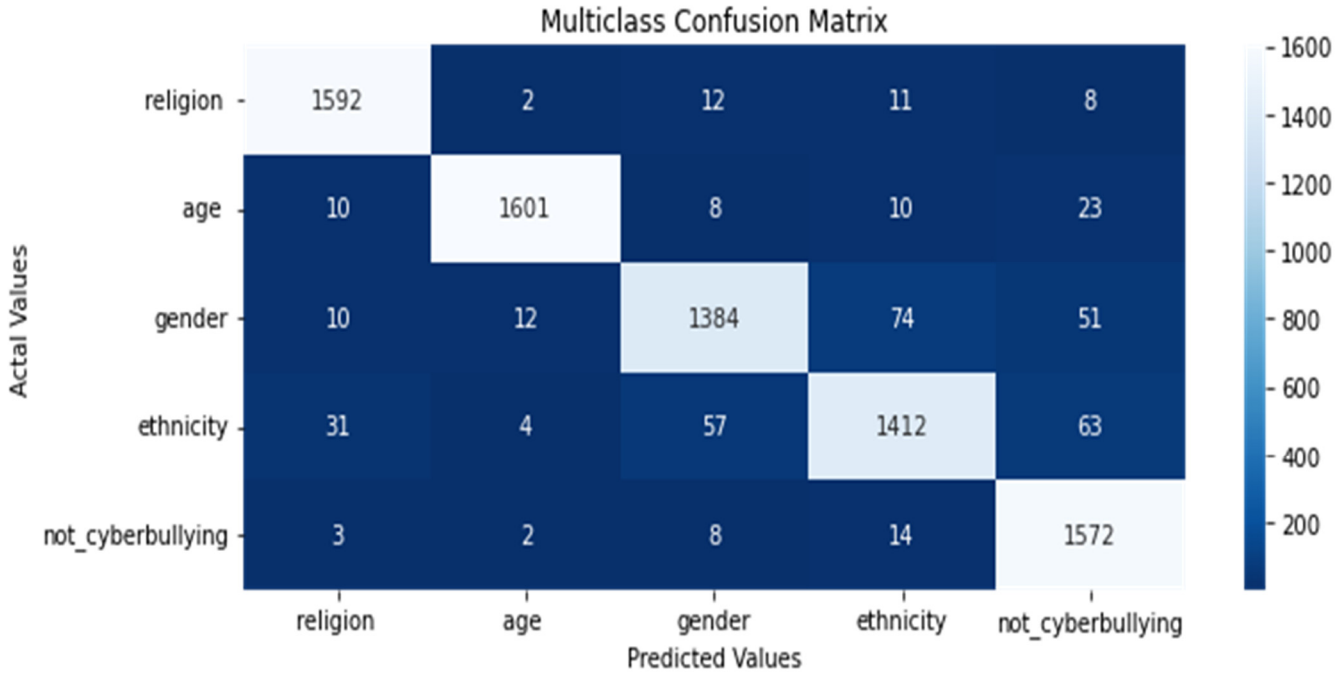

Figure 10. Confusion matrix of the CNN-BiLSTM model using the multiclass classification dataset.

As can be seen in the above Figure 8, out of 7974 samples used in the testing set, 116 tweets were misclassified samples using the BiLSTM model for all classes of the dataset.

Comparing the performance of the models with respect to the misclassification rate, the CNN-BiLSTM misclassified 266 tweet samples as illustrated in Figure 9. Thus, the BiLSTM had a higher classification accuracy than the CNN-BiLSTM model. The experimental results for the multiclass classification dataset are presented in Table 4.

Table 4. Results for the BiLSTM and CNN-BiLSTM models using the multiclass dataset.

Model Name	Class Name	Precision%	Recall%	F1-Score%	Accuracy%
BiLSTM	Religion	1.00	97	98	99
	Age	99	99	99	
	Gender	96	99	97	
	Ethnicity	1.00	1.00	1.00	
	Non-bullying	98	99	98	
CNN-BiLSTM	Religion	97	98	97	95
	Age	99	97	98	
	Gender	94	90	92	
	Ethnicity	93	90	91	
	Non-bullying	92	98	95	

To analyze the results of both models, performance plots are used to visualize the training and validation accuracies in each epoch. As shown in Figure 11, the training accuracy of the BiLSTM started at 94% and reached 99%, while the model's validation accuracy improved from 97.5% to 98.5%. The CNN-BiLSTM model has been increased from 86% to 96%.

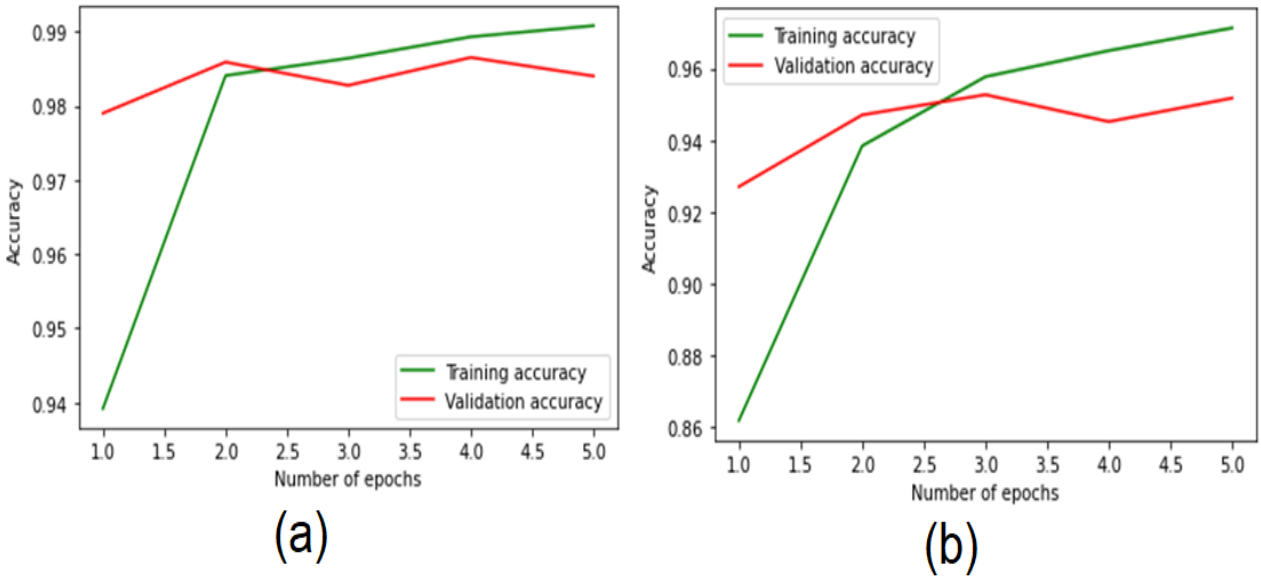

Figure 11. Performance plots o for the multiclass classification dataset (**a**) BiLSTM (**b**) CNN-BiLSTM.

Results from both models were visualized in terms of training and validation accuracies loss at each epoch using performance plots for analysis. As can be seen in Figure 12, the BiLSTM accuracy loss in testing phase is 0.10% where the CNN-BiLSTM is 20%

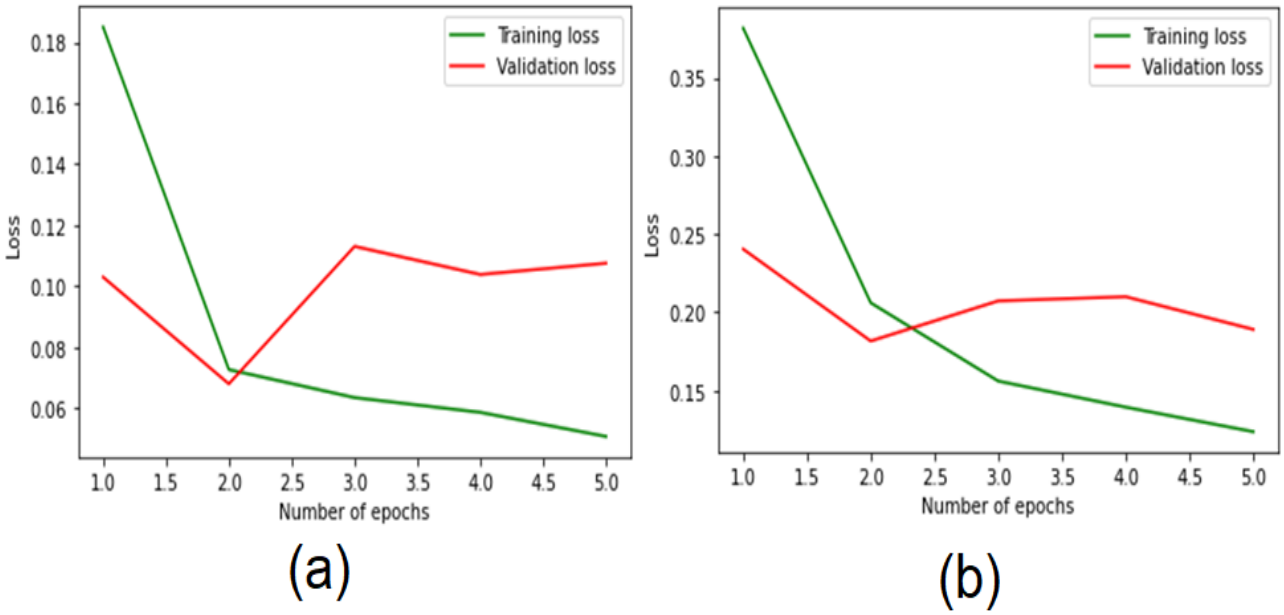

Figure 12. Loss plot for the multiclass classification dataset (**a**) BiLSTM (**b**) CNN-BiLSTM.

4.3. Word Cloud

Word clouds are a technology that is employed in the natural language processing domain to visualize the most significant and frequently used words in a given text. Here, we have used word clouds to visualize the most repeated words in the binary and multiclass cyberbullying classification datasets. Figures 13 and 14 show the word clouds of the two datasets.

(**a**) (**b**)

Figure 13. Word cloud for the binary class dataset (**a**) aggressive and (**b**) non-aggressive and cyberbullying.

Figure 14. Word cloud for the multiclass tweets cyberbullying dataset.

5. Results and Comparison

This subsection presents a comparative analysis of the results of the deep learning models using the multiclass cyberbullying classification dataset along with the GCN and XGBoost approaches presented by Wang et al. [3], using Word2Vec and Keras embedding models and the same dataset. With respect to comparing the binary class dataset classification performance, this is the first research work used this dataset according to the literature review. Table 5 summarizes the results with the same dataset using accuracy as a metric.

Table 5. Comparison the results of the proposed deep learning models with existing methods.

Paper Id	Word Embedding Approach	Technique	Accuracy
Al-Ajlan et al. [27]	Keras embedding	CNN	95%
Hani et al. [43]	TF-IDF	ANN SVM	92.8% 90.3%
Talpur et al. [44]	Word2vec	RF	93%
Wang et al. (2020)	Word2vec	GCN XGBoost	87% 94 %
Proposed work	Keras embedding	BiLSTM CNN-BiLSTM	99% 95

6. Conclusions

Cyberbullying is the use of social media, online discussion blogs, email, or other electronic or digital tools to intimidate, threaten, or coerce others electronically. Also known as abusive digital behavior, cyberbullying is characterized by the use of disparaging, aggressive, or threatening communication. The purpose of this research was to construct and improve a cyberbullying detection system that can be used to analyze and root out instances of online bullying perpetrated by social media users. Deep learning classifiers for detecting hateful online tweets and discussion contents preceding cyberbullying were developed and may be applied in the design of cyberbullying detection systems for online social media sharing platforms, such as Twitter and Facebook. Two different experiments were carried out to train and test the proposed system with binary and multiclass classification datasets. The accuracies of the hybrid deep learning CNN-BiLSTM and single BiLSTM classifiers were compared in the classification of social media posts into dissimilar types of bullying behaviors related to gender, religion, ethnicity, age, aggression, and non-cyberbullying communication. The BiLSM classifier demonstrated promising performance and outperformed the CNN-BiLSM with the binary classification dataset (aggressive or non-aggressive bullying), with a detection rate of 94%. In the case of the multiclass dataset, the BiLSTM is also combined with the CNN-BiLSTM classifier, achieving an accuracy of 99%. The limitations of this study are that the datasets used are limited to the English language and there is overfitting of the proposed models particularly while using the binary class dataset.

For future work, we will focus on developing state of art transformer model for online cyberbullying detection using multilingual datasets.

Author Contributions: Conceptualization, T.H.H.A. and S.N.A.; methodology, M.H.A.-A.; software, M.H.A.-A., T.H.H.A. and S.N.A.; validation, T.H.H.A., M.H.A.-A. and S.N.A.; formal analysis, M.H.A.-A., T.H.H.A. and S.N.A.; investigation M.H.A.-A., T.H.H.A. and S.N.A.; resources, M.H.A.-A., T.H.H.A. and S.N.A.; data curation, S.N.A.; writing—original draft, M.H.A.-A., T.H.H.A. and S.N.A.; preparation, M.H.A.-A., T.H.H.A. and S.N.A.; writing—review and editing, visualization, M.H.A.-A., T.H.H.A. and S.N.A.; supervision, T.H.H.A.; project administration, M.H.A.-A., T.H.H.A. and S.N.A.; funding acquisition, M.H.A.-A., T.H.H.A. and S.N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by the Deanship of Scientific Research at King Faisal University for the financial support under grant No. NA000234.

Data Availability Statement: The data presented in this study are available here: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?select=cyberbullying_tweets.csv (accessed on 18 July 2022).

Acknowledgments: 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 [NA000234].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Englander, E.; Donnerstein, E.; Kowalski, R.; Lin, C.A.; Parti, K. Defining cyberbullying. *Pediatrics* **2017**, *140*, S148–S151. [CrossRef]
2. Johnson, L.D. Counselors and Cyberbullying: Guidelines for Prevention, Intervention, and Counseling. Available online: https://www.counseling.org/docs/default-source/vistas/vistas_2011_article_63.pdf?sfvrsn=f106ccc8_11 (accessed on 18 July 2022).
3. Wang, J.; Fu, K.; Lu, C.T. 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.
4. Slonje, R.; Smith, P.K. Cyberbullying: Another main type of bullying? *Scand. J. Psychol.* **2008**, *49*, 147–154. [CrossRef] [PubMed]
5. Chaffey, D. Global Social Media Research Summary July 2020. Available online: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-globalsocial-media-research (accessed on 20 July 2022).
6. Hosseinmardi, H.; Ghasemianlangroodi, A.; Han, R.; Lv, Q.; Mishra, S. Towards understanding cyberbullying behavior in a semi-anonymous social network. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 17–20 August 2014; pp. 244–252.
7. Cook, S. Cyberbullying Facts and Statistics for 2020. Available online: https://www.comparitech.com/internet-providers/cyberbullying-statistics/ (accessed on 28 July 2022).
8. Yin, D.; Xue, Z.; Hong, L.; Davison, B.D.; Kontostathis, A.; Edwards, L. Detection of harassment on web 2.0. In Proceedings of the Content Analysis in the WEB, Madrid, Spain, 21 April 2009; pp. 1–7.
9. Reynolds, K.; Kontostathis, A.; Edwards, L. Using machine learning to detect cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, Washington, DC, USA, 18–21 December 2011; pp. 241–244.
10. Modha, S.; Majumder, P.; Mandl, T.; Mandalia, C. Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Syst. Appl.* **2020**, *161*, 113725. [CrossRef]
11. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. In Proceedings of Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 6–9 June 2022.
12. Dadvar, M.; Jong, F.D.; Ordelman, R.; Trieschnigg, D. Improved cyberbullying detection using gender information. In Proceedings of the Title of host publicationProceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, 24 February 2012.
13. Kontostathis, A.; Reynolds, K.; Garron, A.; Edwards, L. Detecting cyberbullying: Query terms and techniques. In Proceedings of the 5th Annual Acm, Web Science Conference, online, 2 May 2013; pp. 195–204.
14. Ptaszynski, M.; Masui, F.; Kimura, Y.; Rzepka, R.; Araki, K. Extracting patterns of harmful expressions for cyberbullying detection. In Proceedings of the 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, 27–29 November 2015; pp. 370–375.
15. Zhang, X.; Tong, J.; Vishwamitra, N.; Whittaker, E.; Mazer, J.P.; Kowalski, R.; Hu, H.; Luo, F.; Macbeth, J.; Dillon, E. Cyberbullying detection with a pronunciation based convolutional neural network. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 740–745.
16. Chavan, V.S.; Shylaja, S. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 10–13 August 2015; pp. 2354–2358.
17. Squicciarini, A.; Rajtmajer, S.; Liu, Y.; Griffin, C. Identification and characterization of cyberbullying dynamics in an online social network. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 25–28 August 2015; pp. 280–285.
18. Ozel, S.A.; Saraç, E.; Akdemir, S.; Aksu, H. Detection of cyberbullying on social media messages in turkish. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 366–370.
19. Bozyigit, A.; Utku, S.; Nasiboglu, E. Cyberbullying detection by using artificial neural network models. In Proceedings of the 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 11–15 September 2019; pp. 520–524.
20. Bozyiğit, A.; Utku, S.; Nasibov, E. Cyberbullying detection: Utilizing social media features. *Expert Syst. Appl.* **2021**, *179*, 115001. [CrossRef]
21. Kumari, K.; Singh, J.P.; Dwivedi, Y.K.; Rana, N.P. Towards cyberbullying-free social media in smart cities: A unified multi-modal approach. *Soft Comput.* **2020**, *24*, 11059–11070. [CrossRef]
22. Çiğdem, A.; Çürük, E.; Eşsiz, E.S. Automatic detection of cyberbullying in formspring.me, myspace and Youtube social networks. *Turk. J. Eng.* **2019**, *3*, 168–178.
23. Gomez, R.; Gibert, J.; Gomez, L.; Karatzas, D. Exploring hate speech detection in multimodal publications. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 1470–1478.
24. Kumari, K.; Singh, J.P.; Dwivedi, Y.K.; Rana, N.P. Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. *Future Gener. Comput. Syst.* **2021**, *118*, 187–197. [CrossRef]
25. Sadiq, S.; Mehmood, A.; Ullah, S.; Ahmad, M.; Choi, G.S.; On, B.-W. Aggression detection through deep neural model on Twitter. *Future Gener. Comput. Syst.* **2021**, *114*, 120–129. [CrossRef]

26. Hosseinmardi, H.; Rafiq, R.I.; Han, R.; Lv, Q.; Mishra, S. Prediction of cyberbullying incidents in a media-based social network. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 186–192.
27. Al-Ajlan, M.A.; Ykhlef, M. Deep learning algorithm for cyberbullying detection. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 199–205. [CrossRef]
28. Aldhyani, T.H.H.; Alsubari, S.N.; Alshebami, A.S.; Alkahtani, H.; Ahmed, Z.A.T. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12635. [CrossRef]
29. Pawar, R.; Raje, R.R. Multilingual cyberbullying detection system. In Proceedings of the 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 20–22 May 2019; pp. 40–44.
30. Dataset. Available online: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset (accessed on 12 May 2022).
31. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
32. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
33. Al-Hashedi, M.; Soon, L.K.; Goh, H.N. Cyberbullying detection using deep learning and word embeddings: An empirical study. In Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, Bangkok, Thailand, 23–25 November 2019; pp. 17–21.
34. Zhang, Z.; Robinson, D.; Tepper, J. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In Proceedings of the European semantic web conference, Anissaras, Crete, Greece, 3–7 June 2018; Springer: Cham, Switzerland, 2018; pp. 745–760.
35. Dessì, D.; Recupero, D.; Sack, H. An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments. *Electronics* **2021**, *10*, 779. [CrossRef]
36. Chollet, F. "Keras", GitHub. Available online: https://github.com/fchollet/keras (accessed on 11 August 2022).
37. Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.; Carvalho, J.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic cyberbullying detection: A systematic review. *Comput. Hum. Behav.* **2018**, *93*, 333–345. [CrossRef]
38. Arshi, S.; Zhang, L.; Strachan, R. Prediction using LSTM networks. In Proceedings of the 2019 International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019; pp. 1–8.
39. Alsubari, S.N.; Deshmukh, S.N.; Al-Adhaileh, M.H.; Alsaade, F.W.; Aldhyani, T.H. Development of Integrated Neural Network Model for Identification of Fake Reviews in E-Commerce Using Multidomain Datasets. *Appl. Bionics Biomech.* **2021**, *11*, 5522572. [CrossRef] [PubMed]
40. Understanding LSTM Cells Using C#. Available online: https://msdn.microsoft.com/en-us/magazine/mt846470.aspx (accessed on 16 June 2022).
41. Alzahrani, M.E.; Aldhyani, T.H.; Alsubari, S.N.; Althobaiti, M.M.; Fahad, A. Developing an Intelligent System with Deep Learning Algorithms for Sentiment Analysis of E-Commerce Product Reviews. *Comput. Intell. Neurosci.* **2022**, *10*, 3840071. [CrossRef] [PubMed]
42. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. *arXiv* **2014**, arXiv:1404.2188.
43. Hani, J.; Nashaat, M.; Ahmed, M.; Emad, Z.; Amer, E. Ammar Mohammed. Social media cyberbullying detection using machine learning. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 5.
44. Talpur, B.A.; Declan, O.S. Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in twitter. *Informatics* **2020**, *7*, 52. [CrossRef]