Ingeniería Solidaria

Comparative analysis on deep neural network models for detection of cyberbullying on Social Media

Análisis comparativo sobre modelos de redes neuronales profundas para la detección de ciberbullying en redes sociales

Análise comparativa de modelos de redes neurais profundas para detecção de cyberbullying em redes sociais

Sivadi Balakrishna[1]
Yerrakula Gopi[2]
Vijender Kumar Solanki[3]

Received: October 15th, 2021
Accepted: December 15th, 2021
Available: January 11th, 2022

How to cite this article:
S. Balakrishna, Y. Gopi, V.K. Solanki, “Comparative Analysis on Deep Neural Network Models for Detection of Cyberbullying on Social Media,” *Revista Ingeniería Solidaria,* vol. 18, no. 1, 2022.
doi: https://doi.org/10.16925/2357-6014.2022.01.05

Research article. https://doi.org/10.16925/2357-6014.2022.01.05

[1] Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology & Research (Deemed to be University), Vadlamudi, Guntur, A.P., India.
Email: drsivadibalakrishna@gmail.com
ORCID: https://orcid.org/0000-0002-8939-9307

[2] Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology & Research (Deemed to be University), Vadlamudi, Guntur, A.P., India.
Email: ygopi091@gmail.com
ORCID: https://orcid.org/0000-0002-6047-3825

[3] Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, India.
Email: spesinfo@yahoo.com
ORCID: https://orcid.org/0000-0001-5784-1052

Abstract

Introduction: Social media usage has been increased and it consists of both positive and negative effects. By considering the misuse of social media platforms through various cyberbullying methods like stalking and harassment, there should be preventive methods to control these and avoid mental stress.

*Problem***:** These extra words will expand the size of the vocabulary and influence the performance of the algorithm.

Objective: To detect cyberbullying in social media.

Methodology: In this paper, we come up with variant deep learning models like Long Short Term Memory (LSTM), Bi-Directional Long Short Term Memory (BI-LSTM), Recurrent Neural Networks (RNN), Bi-Directional Recurrent Neural Networks (BI-RNN), Gated Recurrent Unit (GRU), and Bi-Directional Gated Recurrent Unit (BI-GRU) to detect cyberbullying in social media.

Results: The proposed mechanism has been performed, analyzed and implemented on Twitter data with Accuracy, Precision, Recall, and F-Score as measures. The deep learning models such as LSTM, BI-LSTM, RNN, BI-RNN, GRU, and BI-GRU are applied on Twitter to public comments data and performance was observed for these models, obtaining an improved accuracy of 90.4%.

Conclusions: The results indicate that the proposed mechanism is efficient when compared with the state of the art schemes.

Originality: Applying deep learning models to perform comparative analysis on social media data is the first approach to detecting cyberbullying.

Restrictions: These models are applied only on textual data comments. Own work have not concentrated on multimedia data such as Audio, Video, and Images.

Keywords: social media, deep learning, cyberbullying detection, LSTM, GRU, RNN, Bi-directional LSTM, Bi-directional GRU, Bi-directional RNN, Tokenization.

Resumen

Introducción: el uso de las redes sociales se ha incrementado y tiene efectos tanto positivos como negativos. Al considerar el uso indebido de las plataformas de redes sociales a través de varios métodos de acoso cibernético, como el acecho y el acoso, debe haber métodos preventivos para controlarlos y evitar el estrés mental.

Problema: estas palabras adicionales ampliarán el tamaño del vocabulario e influirán en el rendimiento del algoritmo.

Objetivo: Detectar el ciberacoso en las redes sociales.

Metodología: en este documento, presentamos variantes de modelos de aprendizaje profundo como la memoria a largo plazo (LSTM), memoria bidireccional a largo plazo (BI-LSTM), redes neuronales recurrentes (RNN), redes neuronales recurrentes bidireccionales (BI-RNN), unidad recurrente cerrada (GRU) y unidad recurrente cerrada bidireccional (BI-GRU) para detectar el ciberacoso en las redes sociales.

Resultados: El mecanismo propuesto ha sido realizado, analizado e implementado sobre datos de Twitter con Accuracy, Precision, Recall y F-Score como medidas. Los modelos de aprendizaje profundo como LSTM, BI-LSTM, RNN, BI-RNN, GRU y BI-GRU se aplican en Twitter a los datos de comentarios públicos y se observó el rendimiento de estos modelos, obteniendo una precisión mejorada del 90,4 %.

Conclusiones: Los resultados indican que el mecanismo propuesto es eficiente en comparación con los esquemas del estado del arte.

Originalidad: la aplicación de modelos de aprendizaje profundo para realizar un análisis comparativo de los datos de las redes sociales es el primer enfoque para detectar el ciberacoso.

Restricciones: estos modelos se aplican solo en comentarios de datos textuales. El trabajo propio no se ha concentrado en datos multimedia como audio, video e imágenes.

Palabras clave: redes sociales, aprendizaje profundo, detección de ciberacoso, LSTM, GRU, RNN, LSTM bidireccional, GRU bidireccional, RNN bidireccional, tokenización.

Resumo

Introdução: o uso das redes sociais tem aumentado e tem efeitos positivos e negativos. Ao considerar o uso indevido de plataformas de mídia social por meio de vários métodos de cyberbullying, como stalking e bullying, deve haver métodos preventivos para controlá-los e evitar o estresse mental.

Problema: Essas palavras adicionais aumentarão o tamanho do vocabulário e afetarão o desempenho do algoritmo.

Objetivo: Detectar cyberbullying nas redes sociais.

Metodologia: Neste artigo, apresentamos variantes de modelos de aprendizado profundo, como memória de longo prazo (LSTM), memória de longo prazo bidirecional (BI-LSTM), redes neurais recorrentes (RNN), redes neurais recorrentes bidirecionais (BI-RNN), unidade recorrente fechada (GRU) e unidade recorrente fechada bidirecional (BI-GRU) para detectar cyberbullying nas redes sociais.

Resultados: O mecanismo proposto foi executado, analisado e implementado nos dados do Twitter tendo como medidas Accuracy, Precision, Recall e F-Score. Modelos de deep learning como LSTM, BI-LSTM, RNN, BI-RNN, GRU e BI-GRU são aplicados no Twitter a dados de comentários públicos e foi observado o desempenho desses modelos, obtendo uma precisão aprimorada de 90,4%.

Conclusões: Os resultados indicam que o mecanismo proposto é eficiente em comparação com esquemas de última geração.

Originalidade: A aplicação de modelos de deep learning para realizar análises comparativas de dados de redes sociais é a primeira abordagem para detectar cyberbullying.

Restrições: Esses modelos se aplicam apenas a comentários de dados textuais. O trabalho próprio não se concentrou em dados multimídia como áudio, vídeo e imagens.

Palavras-chave: redes sociais, aprendizado profundo, detecção de cyberbullying, LSTM, GRU, RNN, LSTM bidirecional, GRU bidirecional, RNN bidirecional, tokenização.

1. INTRODUCTION

As the growth of digitization has increased rapidly and become part of everyone's life e.g., usage of various platforms of social media like sharing text, audio, videos through Twitter, Facebook, Instagram, and many others using the internet, at the same time, misuse has also grown rapidly, which creates negative impacts on society; misuse often comes in the form of lambasting, which should be hindered. This process of harassing people through offensive comments or sharing others' personal information by electronic means is known as cyberbullying also known as online bullying

[1]. This has become common, especially among teenagers. This bullying includes racism (e.g., skin tone, features), physical appearance (e.g., fat, thin, ugly), sexism (e.g., female, male) and masquerading (creating fake accounts to harass others). Therefore, cyberbullying is hard to detect and trace, which will have intense effects [2]. Therefore, detection of cyberbullying at its primary stage is a pivotal step to eluding any lethal incidents caused by it. However, to minimize or reduce these effects, researchers have developed various machine learning and deep learning models for handling the cyberbullying detection problem [3].

Online social media sites empower users to communicate and display their opinions and feelings openly and anonymously with others. This can occur as an array of tech-empowered exercises, e.g., photo posting, tweeting, social gaming, social video sharing, business networks, feedback & ratings, among many others [4]. The content in these social networks is a rich platform to analyze these feelings and use or misuse. This growing rise in social networking brings ongoing harassment of the so-called "cyberbullying" [5]. Widespread cyberbullying can occur in a variety of ways, including bigotry (e.g., body shaming, skin color), misogyny (e.g., male, female), obesity, intellect (e.g., fat, stupid), and more. This cyberbullying act is often anonymous, i.e., quite harder to pinpoint, and has profound and catastrophic consequences [6]. Therefore, detection at the early point of cyberbullying is a critical step in preventing this and preventing fatal incidents. Different machine learning and deep learning approaches have been used to solve the challenge of cyberbullying in recent years [7]. Another important task for cyberbullying research is to include the appropriate evidence needed to design models to identify cyberbullying. Some datasets for this special task, including the training set in the CAW 2.0 Workshop and the Twitter Bullying Traces dataset, are publicly accessible [8].

As per statistics, over 40 percent of teenagers have become victims of cyberbullying in the United States. In this paper, we have proposed deep learning models for the identification of cyberbullying [9]. We mainly focused on textual cyberbullying, as text is the most common part of social media; like sharing comments, messages, etc. The main requirement for detecting bullying over social media platforms is having accurate data. Social media data will generally be of the heterogeneous form with noise and missing values; various deep learning techniques should be performed on that data to remove noise [10]. In this work, we are dealing with the Twitter dataset i.e., Twitter comments. The task here is to identify offensive and non-offensive comments and improve the accuracy of the results using deep learning models. As deep learning is a new technology and has more advantages than machine learning, so the accuracy of the model will be increased [11-12]. The stress resulting from cyberbullying leads many to suicidal thoughts. The background work will be discussed in the sections below [13].

Machine learning or deep learning algorithms aid researchers in their understanding of big data [14]. There will be a lot of knowledge about people and their cultures in this Big Data era, but it was historically impractical to gather this information [15]. Networking sites are one of the largest repositories of human data. Despite deep learning having comparatively recent applications such as text classification [16], and subject categorization, the effective implementation of deep learning has been observed [17]. There are two main models for deep learning - CNNs and RNNs [18]. Both models use the words embedded in the text series for input and produce vectors for the words that are valued properly. In terms of sentence and question classification, CNNs have been added and proven to the advanced achievement of conventional machine learning techniques such as SVM and MaxEnts [19]. In addition, RNNs are designed to model the text series to increase multi-class learning efficiency [20]. The enhanced RNNs such as LSTMs [21] and Gated Recent Units [22] and Bidirectional LSTMs are used extensively on the NLP application, as they have long-range dependence and can store historical information over time. Their use in NLP applications has been significantly increased in terms of their long-term dependence. Moreover, in various real-time social media implementations, the positive effects of deep learning are rapidly being observed [23]. These include abuse of bigotry and detection of sexism [24], open/covert/non-aggressive analysis of messages, categorization of crisis details (i.e., advice, gifts, infrastructure, compassion) [25], and recognition of domestic violence critical web posts [26].

The following contributions are made to this proposed work to detect cyberbullying in social media.

- To develop a comment-level embedding for this task
- To design a model based on a deep learning early detection framework for cyberbullying detection
- To perform a comparative analysis over LSTM, GRU, RNN, Bidirectional LSTM, Bidirectional GRU, Bidirectional RNN models on the data sets such as Twitter comments and public data comments.
- To evaluate the performance of deep neural networks over Twitter datasets using various measures such as Accuracy, Recall, F-measure, and Precision.

- Implementation of LSTM, GRU, RMN, Bidirectional LSTM, Bidirectional GRU, Bidirectional RMN models on the data set like Twitter comments, public data comments.
- After implementing these models, the accuracy has been improved to 90.2%; which was more than other deep learning models like CNN because the models like RMN that we have implemented are reusable activation functions and can work on huge data sets.

The rest of this paper is organized as follows: Section 2 shows the related work of this study. The proposed methodology for detection of cyberbullying in social media have been described in Section 3. Section 4 shows the performance evaluation of proposed work and it includes the experimental results and analysis. Finally, Section 5 concludes this paper and outlines future enhancements.

2. RELATED WORK

In this section, cyberbullying techniques that efficiently use different models are discussed. Most of the researchers have carried out their work using either machine learning models or deep learning models to detect cyberbullying. Semiu salawu et al. [27] have introduced automated cyberbullying detection using features such as spelling, document length, capitalization for detection tasks on data sets to spring. After obtaining the dataset, the pre-processing measures are taken care of, like tokenization and stemming. Classifiers such as Naive Bayes, J48, and SVM are used to determine cyberbullying using binary classification i.e., whether data is bullying or non-bullying. Rui Zhao et al. [28] proposed cyberbullying detection based on a semantic-enhanced marginalized denoising auto encoder. This model was developed using a deep learning technique, namely semantic enhanced marginalized denoising autoencoder (smSDA); it was the extension for the SDA deep learning model and acheived approximately 69% accuracy after performing these techniques on the Twitter and Myspace datasets. The main advantage of smSDA was in being able to more robustly determine discriminative messages, with word embeddings used to expand the bullying word list. Lu Cheng et al. [29] have developed unsupervised cyberbullying detection via a time-informed Gaussian mixture model. Because of the constraints of the time-consuming labeling processes and as labels may not be generalized for future use due to variant languages, they opted for unsupervised learning. This model enciphers social media by various features like network, time, text, and estimates bullying based on the Gaussian mixture model. The datasets they experimented with are Instagram and Vine, achieving an accuracy of 87%. Kirti Kumari et al. [30] worked for cyberbullying-free social media in smart cities using a multi-modal approach; they aimed at identifying text bullying as well as images. The single-layer convolution neural network model has been used for improving performance data from popular platforms like those that Facebook, Instagram, and Twitter. After obtaining data from these popular sites, they are classified as either text or images and the TF-IDF technique and CNN are applied for determining bullying over the data. This model achieved an accuracy of 78%.

Mohammed Ali et al. [31] developed cyberbullying prediction methods on social media in the big data era using machine-learning techniques. The text classification approach was used for identifying bullying instead of lexical-based classification because, in the lexical analysis phase, the text or comments may be in unordered form making it difficult to identify; resulting in inaccurate results. Therefore, different machine learning techniques like SVM, KNN, RF, and DT were applied on data sets of social media platforms like Twitter, Instagram. Akshi Kumar et al. [32] proposed multimodal cyberbullying detection using capsule network with dynamic routing. They developed cyberbullying detection methods using a deep neural model with different modalities of data, like text and visual. A capsule dynamic routing model was used to determine textual bullying, whereas a deep convolution neural network was used in the case of visual bullying prediction. These models are applied to data obtained from YouTube, Instagram, and Twitter comments and achieved an AUC-ROC of 0.98. Paul Sayanta et al. [33] have developed a deep learning-based multi-modal approach for cyberbullying identification. As many works employ unimodal approaches, they worked on multimodal approaches like LSTM for text classification and RCNN for visual classification and achieved an F-score of 0.75. These techniques have been applied to Vine data. Nevertheless, the achieved accuracy was not satisfiable when compared with deep learning models. Jyoti Prakash Singh et al. [34] implemented identification of cyberbullying on social media using a genetic algorithm. As uni-modal approaches will not give results that are more accurate, they implemented a multi-model approach like VGG-16 for text classification and CNN for visual classification. For improving the accuracy, they applied a genetic algorithm and obtained an F-score of 78%. The dataset used was Twitter and Instagram. However, the drawback was that the developed model was not accurate on multi-lingual comments. Amgad Munner et al. [35] proposed detecting cyberbullying on the social media platform of Twitter using a comparative study of machine learning algorithms like LR, SVM, NB, RF. The results obtained from these techniques demonstrated that LR achieved an accuracy of around 90%.

Bandeh Ali Talpur et al. [36] implemented a machine learning model for the detection of cyberbullying of various features like lexicon; machine learning algorithms like KNN, Decision tree, Random forest are applied to the Twitter data. After applying these techniques, the results are compared and concluded; among these, Decision tree models obtained the greatest accuracy. Gautam Srivastava et al. [37] proposed deep learning models for determining cyberbullying among social media platforms. Variants of deep learning models like BLSTM, GRU, RNN are implemented on social media platforms like Instagram and Twitter; the results concluded that, among all these models, the B-LSTM model achieved more accuracy in detecting cyberbullying. Nijia Lu et al. [38] performed cyberbullying detection in social media based on character level convolution neural networks. They worked based on textual information because the text is the most common form of social media data. The data set used was Chinese Sina Weibo data. On this data set, the character level CNN technique was used. Therefore, in this model, characters are the smallest learning units for detecting bullying. After applying these techniques, the obtained F-score was 71. Sriparna Saha et al. [39] have developed cyber BERT for cyberbullying identification. Deep learning techniques are used on different social media platforms like Formspring, Twitter, and Instagram. For variant language, BERT (bidirectional encoder representations from transformers) have been used on the obtained data set; variant deep learning techniques like LR, CNN, and LSTM have been applied and achieved an F-score of 0.83. Cynthia van hee et al. [40] have proposed automatic cyberbullying detection among social media platforms. By performing binary classification using linear vector support machines on different social media platforms like Twitter and Instagram they obtained an accuracy of 64%. This achieved accuracy was not up to standard and can be improved using deep learning models. David van Bruwaene et al. [41] have developed a multi-platform dataset for detecting cyberbullying among social media. As most of the researchers performed their work on single social media platforms, they have chosen multi-data sets and applied variant machine learning algorithms along with a multi-technique annotation system. The data sets chosen are YouTube and Twitter; an accuracy of about 85% was achieved. Hugo Rosa et al. [42] proposed cyberbullying detection among social media platforms. Instead of using machine-learning techniques, they have chosen deep learning models like CNN, LSTM. The data sets chosen to perform these analyses are google –news, Twitter and Formspring. After applying these techniques, they achieved an F-measure of about 0.82. Akshi Kumar et al. [43] developed cyberbullying detection among social media using soft-computing techniques. Meta-Analysis has been performed on the data sets chosen and then applied with machine learning techniques. They have chosen various social media platforms' data like Twitter and Facebook. After identifying cyberbullying among data sets, they are further classified into ICB (indirect bullying) or DCB (direct bullying). Maral Dadvar et al. [44] proposed cyberbullying detection among social media platforms using deep learning models. They performed the analysis by applying variant deep learning models on Wikipedia, Twitter and Formspring; after they also expanded their work on the YouTube data set. The deep learning techniques applied are CNN, LSTM. They have obtained better accuracy comparatively with machine learning models.

Table 1. Survey of the various techniques used for detection of cyberbullying

Reference	Word Embedding Technique	Approach	Algorithms/ Techniques used	Datasets
[27]	TF	Machine Learning	Naive Bayes, j48, SVM	Formspring
[28]	TF-IDF	Deep Learning	smSDA	Twitter
[29]	Binary Encoding	Unsupervised Learning	Gaussian mixture model	Instagram, Vine
[30]	Latent semantic analysis	Multi-model Approach	TF-IDF, CNN	Twitter, Instagram
[31]	Word2Vec Embedding	Machine Learning	KNN, RF	Instagram, Twitter
[32]	Latent semantic analysis	Multi-model Approach	CNN	YouTube, Instagram
[33]	TF	Deep Learning	LSTM, RCNN	Vine
[34]	Word2Vec Embedding	Multi-model Approach	VGC-16, CNN	Instagram, Twitter
[35]	Latent semantic analysis	Machine Learning	LR, NB, RF, SBD	Twitter
[36]	TF-IDF	Machine Learning	KNN, Decision Tree, RF	Twitter
[37]	Binary Encoding	Deep Learning	BLSTM, RNN	Instagram, Twitter
[38]	Word2Vec Embedding	Deep learning	Character level CNN	Chinese Weibo
[39]	Latent semantic analysis	Deep Learning	LR, CNN, LSTM	Twitter, Formspring
[40]	TF	Machine Learning	Linear Vector Support Machine	Twitter, Instagram
[41]	Word2Vec Embedding	Machine Learning	Multi-Technique Annotation	YouTube, Twitter
[42]	Latent semantic analysis	Deep Learning	CNN, LSTM	Google-news, Twitter
[43]	TF-IDF	Machine Learning	NLP	Twitter, Facebook
[44]	Binary Encoding	Deep Learning	CNN, LSTM	Wikipedia, Formspring

Source: Own work

Table 1 shows the survey of the various techniques used for the detection of cyberbullying. The following limitations are observed over the recent studies on the detection of cyberbullying in social media.

- After analyzing the literature survey, we observed that most of the researchers have performed their work using machine-learning models.
- Nevertheless, machine-learning models are less accurate when compared with deep learning models; this being the drawback of most of the works analyzed here.
- To overcome the drawback, we have implemented deep learning techniques to improve accuracy.
- The majority of the DL-based developed models tend to work more accurately on small data.
- Most of the researchers who have implemented deep learning techniques used LSTM and CNN techniques only.

3. PROPOSED METHODOLOGY

We have proposed an architecture by comparing various deep neural networks. We have taken six deep learning models LSTM, GRU, RNN, BI-LSTM, BI-GRU, BI-RNN. We have compared the accuracies, and we have found which model works best. We got an accuracy of 90.4 for Bi-directional GRU, which is more than the remaining models. Fig. 1 depicts the proposed work flow for detection of cyberbullying in Twitter data.

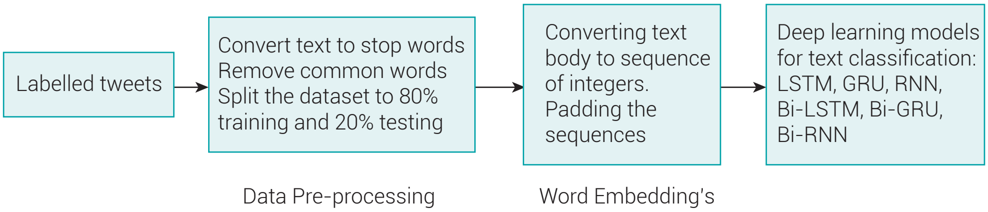
Labelled tweets
Convert text to stop words
Remove common words
Split the dataset to 80% training and 20% testing
Converting text body to sequence of integers.
Padding the sequences
Deep learning models for text classification: LSTM, GRU, RNN, Bi-LSTM, Bi-GRU, Bi-RNN
Data Pre-processing
Word Embedding's

Fig.1 Work flow of the proposed work
Source: Own work

We can identify the most commonly occurring words such as bitch, pussy, ass, and so on, in very high numbers. These words tell us what spam messages look like and their features. We separate the dataset into training and testing sets, and mostly the data is used for training. Using similar data for testing and training helps to understand the characteristics of the model.

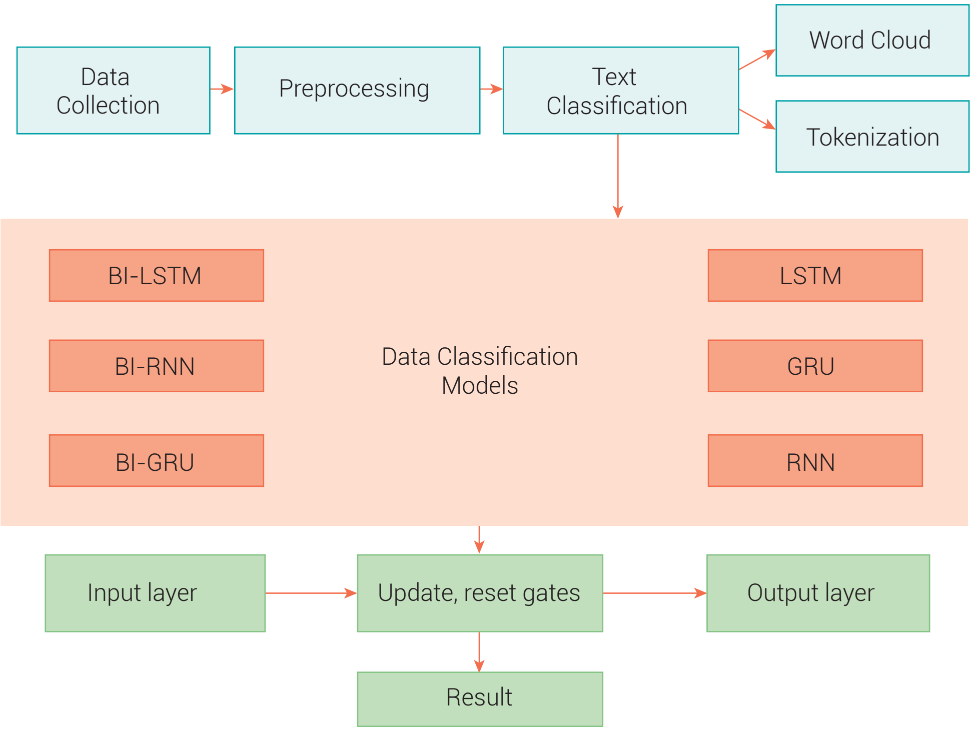

Fig. 2 Proposed Architecture for detection of cyberbullying

Source: Own work

The detailed overview of the proposed methodology is shown in Fig.2 and described as follows:

i) **Dataset collection:** The first step in the model includes the collection of datasets as labelled tweets.

ii) **Data Pre-processing:** The second step does the pre-processing that includes handing the missing values, dimensionality reduction, and vectorization, which is used to speed up the computation.

- **Splitting the data:** The data is divided into a training set and testing set. By default, the ratio of training to testing split is 80:20 (percent) respectively. i.e., 0.8 data is taken into training and 0.2 is taken into testing.

iii) **Tokenization:** Converting the text into a sequence of integers or vectors and all punctuations are removed. These sequences are then split into lists of tokens. Later, they will then be indexed or vectorized.

iv) **Evaluation of the model:** Once the training is done, the model is evaluated and the testing set is evaluated later based on how the model is trained.

The performance can be measured using performance metrics that were discussed in the Experiment Evaluation section.

v) **Deep Learning algorithms:** The algorithms that were used for training our model are discussed in detail below.

3.1 DEEP LEARNING MODELS

In this section, the various Deep Learning models that are used for comparative analysis on the detection of cyberbullying are discussed.

3.1.1 LSTM

The LSTM model is a special kind of RNN. Memorizing information for a long time is their default behavior, thus avoiding long-term dependency problems. LSTM consists of a chain-like structure and the module, which repeats, will have a difference in structure to RNN. Here, LSTM does not change information but the gates do; they regulate the information. Gates consist of multiplication operations and a sigmoid neural net layer. This sigmoid layer generally provides two outputs like 0 and 1. Here 0 means "will not allow any information" and 1 means "it allows information". LSTM consists of three gates.

Here the first sigmoid layer, called the "forget gate layer", takes *ht*–1 and *xt*

- The next most important decision is to identify what new information we should store in cell state. This again consists of two steps:
    1. Input gate layer (It decides what values should be updated)
    2. Tanh layer (It creates new vector values)

Finally, we should combine the above two steps.

- Now, we should update the old cell state into a new state.
- Then we complete the sigmoid layer; that will decide on what we are going to put as the output.
- After that, we run the Tanh layer and multiply this with the output obtained from the sigmoid gate.

3.1.2 BIDIRECTIONAL LSTM

This was the extension technology of LSTM. It is given the name "bidirectional" as it can train two inputs instead of one; first is the input sequence, and then the reverse of it. However, the working of a Bidirectional LSTM is similar to that of an LSTM with the main differences being:

- In the case of Bidirectional LSTM, we have forward LSTM and Backward LSTM, so that we can easily predict the necessary result.
- Nevertheless, Bidirectional LSTM is complex when compared to LSTM.
- As observed from the diagram, the results obtained from the backward layer by various LSTM implementations and forward layer are merged.

3.1.3 GRU

This was the newer generation of RNN and was similar to LSTM. Unlike LSTM it doesn't have cell states; instead, it uses the hidden state to transfer the information. It consists of only two gates, i.e., reset gate and update gate. The following algorithm is used for the detection of cyberbullying with GRU model.

Step-1: Take the dataset
$\sum_{i=1}^{n} X_i$ Where n=24,783 samples and x_i input features
Step-2: Convert the text to stop words using word cloud
Step-3: Convert the text to vectors to send the text to a neural network using tokenization
Step-4: Send this converted vector to a Bi-GRU deep neural network and calculate validation and testing accuracies.

$$R_t = \rho\ (X_t\ W_{xr} + H_{t\text{-}1}\ W_{hr} + b_r\ )$$
$$Z_t = \rho\ (X_t\ W_{xz} + H_{t\text{-}1}\ W_{hz} + b_z\ )$$
$$H_t = \tanh\ (X_t\ W_{xh} + (R_t\ H_{t\text{-}1})\ W_{hh} + b_h\ )$$
$$H_t = Z_t\ H_{t\text{-}1} + (1\text{-}Z_t\ )\ H_t$$

Step-5: Find the Precision, Recall, and F1 Score.
Step-6: Plot the confusion matrices for every model and compare the results.

In addition to that, the following steps are required to update and reset the gates for adjusting the weights.

1. Starting with update gate

$$Z_t = \rho \ (W^{(z)} x_t + U^{(z)} h_{\ t\text{-}1}$$

Here x_t is multiplied with its weight w(z). Similarly, h_(t-1) is multiplied with its weight U(z).

2. Reset gate equation

$$r_t = \rho \ (W^{(r)} x_t + U^{(r)} h_{\ t\text{-}1}$$

This was similar to that of update gates; the difference comes with weights and gates usage.

3.1.4 BIDIRECTIONAL-GRU

This consists of two GRUs. One is for taking input in the forward direction and another for taking the input in the backward direction. It consists of only input and forget gates. Fig.3 shows the basic Bi-GRU model in a detailed way.

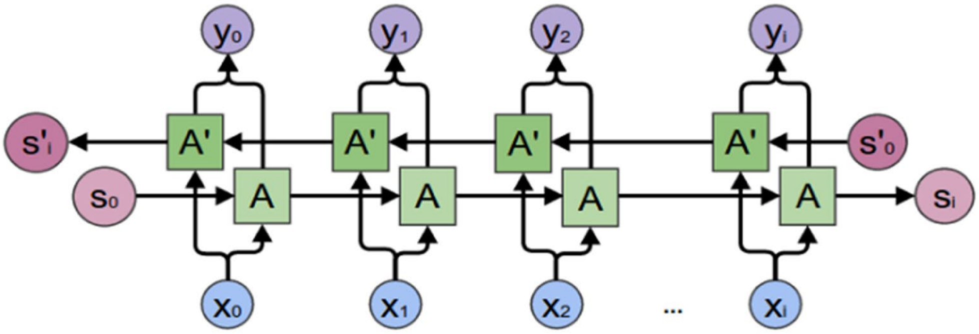

Fig. 3 Bi-GRU model
Source: Own work

Consequently, every A and A' can be replaced with the working Bi-GRU model that has been deliberated in Fig.4 below.

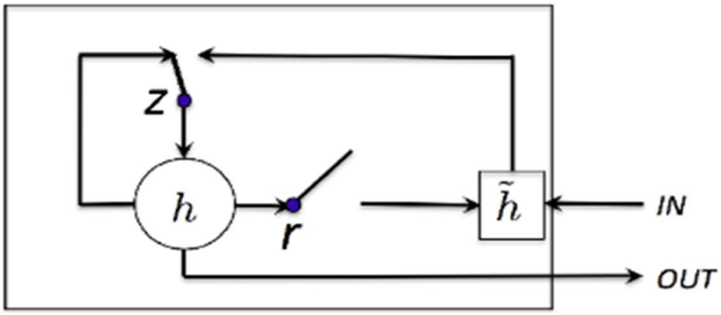

Fig. 4 Bi-GRU working model
Source: Own work

Here, Z is the update gate and r is the reset gate.

3.1.5 RNN

RNN is termed as recurrence neural networks; these are developed to work on sequential problems. Normal ANN cannot handle sequential data as they do not have memory concepts. RNN consists of multiple ANN. Therefore, RNN can work on image processing, sentiment classification etc. Fig.5 shows the RNN architectural model.

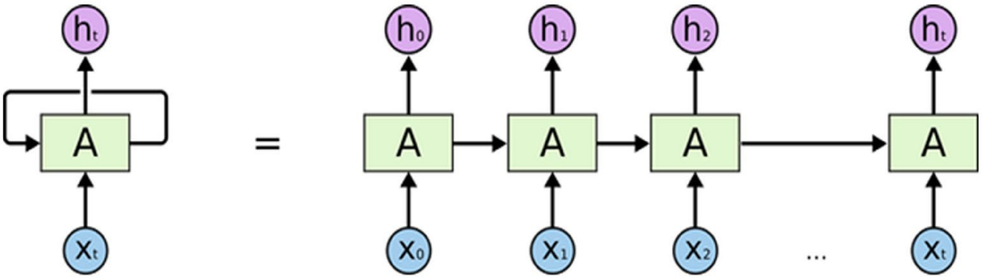

Fig. 5 RNN model
Source: Own work

3.1.6 BI-DIRECTIONAL RNN

BI-RNN is similar to that of RNN. It consists of regular RNN in two directions. One is for the forward direction and another for the backward direction. Fig.6 shows the Bi-RNN model used for the detection for cyberbullying and used for the comparative analysis.

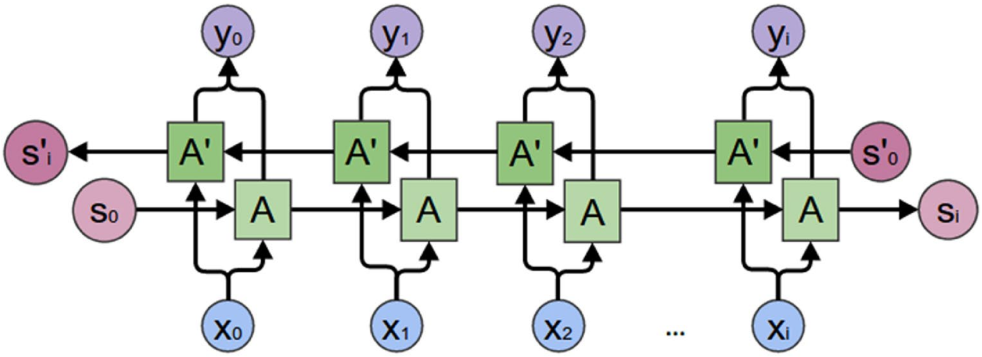

Fig. 6 Bi-RNN model
Source: Own work

Here, Bi-RNN is used to compare the output and input as they can work in reverse order; by that, we can check the errors and accuracy.

4. PERFORMANCE EVALUATION

The proposed work has been implemented on Anaconda navigator (Jupiter notebook). This was programmed using Python code with a DELL laptop of I7 Intel Pentium Processor with GPU support, 8 GB RAM, and 1 TB HDD on the Windows 10 platform.

4.1 Dataset details

We evaluated the Deep Learning models taken on two datasets. The first dataset is the Twitter dataset that consists of 11091 rows and 2 columns. It is a labelled binary classifier dataset. The labels here are offensive and non-offensive. The second dataset corresponds to the Public comments dataset. This Dataset consists of 24784 rows and 2 columns. It is also a labelled binary classifier dataset. The labels here also are offensive and non-offensive. Table 2 shows the dataset details used for testing the deep learning models.

Table 2. Dataset details

Dataset Name	Instances	Features	Category	URL
Twitter Dataset	11091	2	Offensive, Non-Offensive	(https://github.com/dhavalpotdar/cyberbullyingdetection/blob/master/labeled tweets.csv)
Public dataset	24784	2	Offensive, Non-Offensive	(https://raw.githubusercontent.com/dhavalpotdar/cyberbullyingdetection/master/public_data_labeled.csv)

Source: Own work

4.2 Performance Metrics

To evaluate the performance of the proposed work, the following metrics are considered for measuring the proposed work. These metrics are generated from the confusion matrix as shown in Fig 7.

	Predicted as "YES"	**Predicted as "NO"**
Actually as "YES"	True Positive $[B_x \rightarrow B_x]$	False Negative $[B_x \rightarrow NB_x]$
Actually as "NO"	False Positive $[NB_x \rightarrow B_x]$	True Negative $[NB_x \rightarrow NB_x]$

Fig.7 Confusion Matrix
Source: Own work

4.2.1. True positive

$B_x \rightarrow B_x$: This is an estimation of detected bullying words considered correctly as detected bullying words.

4.2.2. True negative

$NB_x \rightarrow NB_x$: This is an estimation of non-detected bullying words considered correctly as non-detected bullying words.

4.2.3. False positive

$NB_x \rightarrow B_x$: This is an estimation of non-detected bullying words considered incorrectly as detected bullying words.

4.2.4. False negative

$B_x \rightarrow NB_x$: This is an estimation of detected bullying words considered incorrectly as non-detected bullying words.

4.2.5. Accuracy

Accuracy is the first step towards performance measure where it defines the ratio between the 'total count of correctly detected bullying words' and the 'total count of detected bullying words' as shown Eq.1.

$$Accuracy = \frac{(B_x \rightarrow B_x + NB_x \rightarrow NB_x)}{[B_x \rightarrow B_x + NB_x \rightarrow NB_x + NB_x \rightarrow B_x + B_x \rightarrow NB_x]} \quad (1)$$

4.2.6. Precision, Recall & F-measure

Precision relates to the exactness of the data, and the Recall about completeness. Combined, precision and recall conclude the accuracy of the system, whereas the accuracy does not explain much about the false results. F-measure studies precision and recall to decide upon the score. Its harmonic mean over precision and recall as shown Eq.2-4.

$$Precision = \frac{B_x \rightarrow B_x}{[B_x \rightarrow B_x + NB_x \rightarrow B_x]} \quad (2)$$

$$Recall = \frac{B_x \rightarrow NB_x}{[B_x \rightarrow B_x + B_x \rightarrow NB_x]} \quad (3)$$

$$F - measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \quad (4)$$

4.3 Result Discussions

In this section, we are going to analyze various deep learning models with respect to several performance evaluation metrics.

Table 3. Results obtained for various deep learning models

Deep learning models	Precision	Recall	F1-score	Accuracy
RNN	0.88	0.89	0.87	89.8
LSTM	0.90	0.90	0.90	90.09
GRU	0.89	0.89	0.89	89.4
BIRNN	0.89	0.90	0.89	89.5
BILSTM	**0.90**	**0.90**	**0.90**	**90.21**
BIGRU	0.89	0.90	0.89	90.0

Source: Own work

Table 3 represents the results obtained after applying various deep-learning techniques like RNN, BI-RNN, LSTM, BI-LSTM, GRU, and BI-GRU over a public data set. After applying these techniques, accuracy, precision, recall and F1-score were calculated using the confusion matrix obtained. We concluded that the BI-LSTM method has acquired more accuracy.

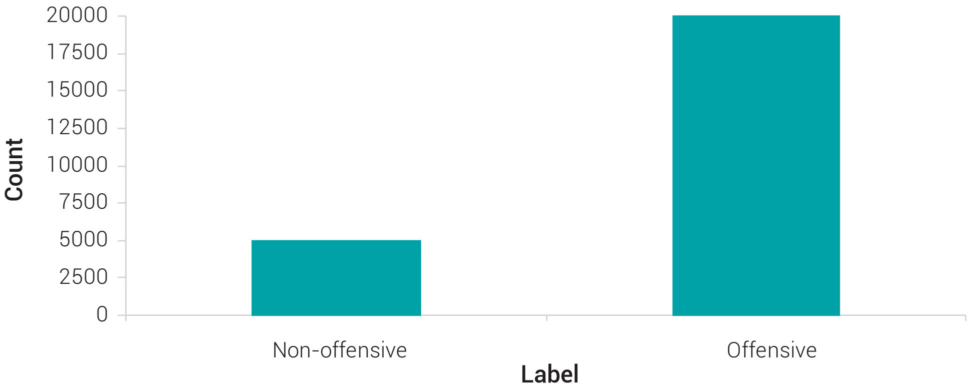

Fig.8 Count of offensive and non-offensive labels of dataset1
Source: Own work

Fig.8 depicts the number of offensive and non-offensive comments that are obtained from dataset1. So, from the diagram we can depict that there were over 5000 Non-offensive comments and 20,000 offensive comments from public dataset.

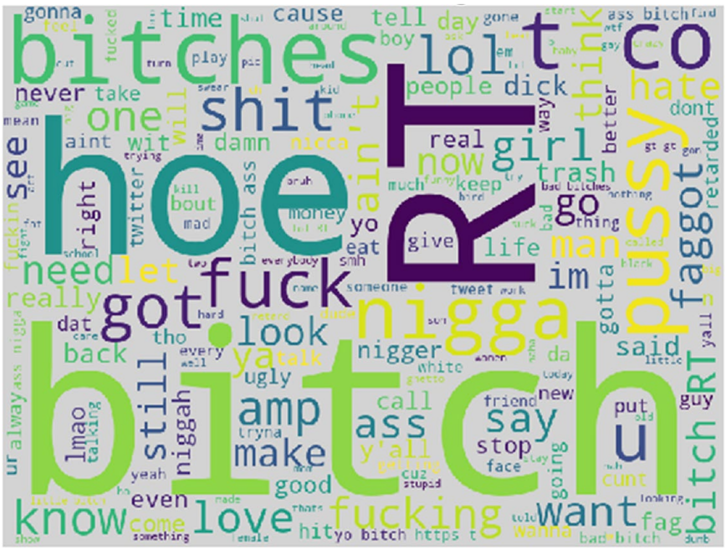

Fig.9 Offensive messages
Source: Own work

Fig.9 represents the word cloud of offensive messages which are obtained from the public dataset. Some of the offensive messages are amp, ass, bitch etc. The primary aim of the project was to find out the offensive and non-offensive comments from the data set.

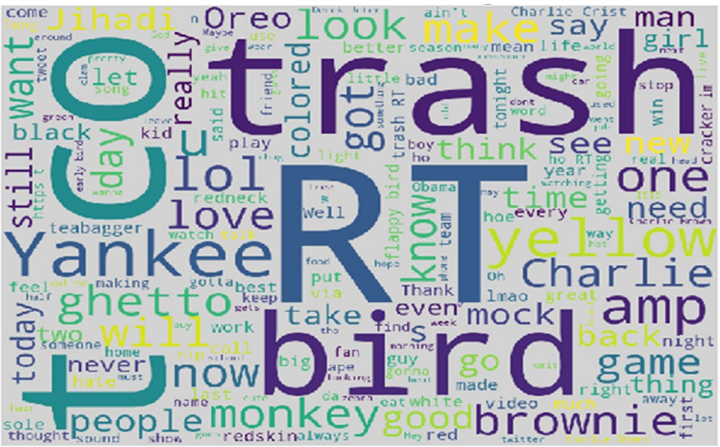

Fig.10 Non-offensive messages
Source: Own work

Fig.10 represents the Non-offensive messages. Some of the non-offensive messages are trash, bird, love etc. The primary aim of the work was to find out the offensive and non-offensive comments from the data set.

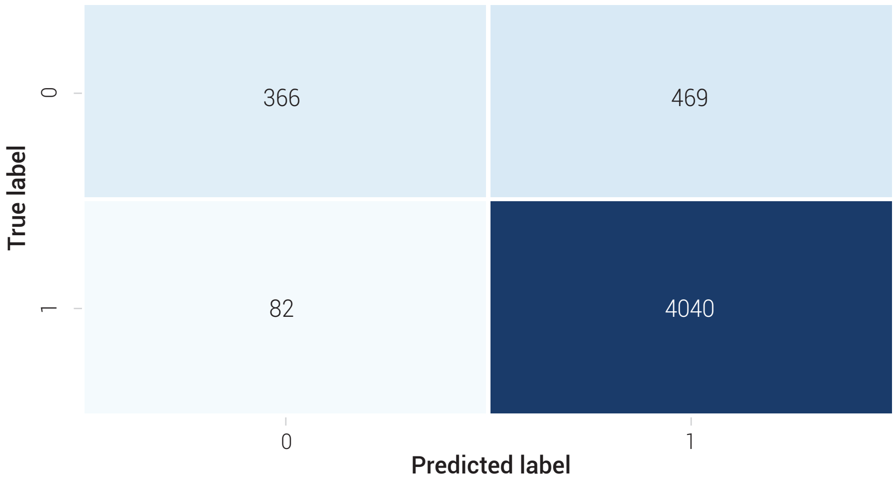

Fig.11 RNN model confusion matrix
Source: Own work

Fig.11 represents the confusion matrix that was obtained after applying the RNN model on dataset1. This matrix was obtained after applying the RNN method on the public data set. The TN=366, FP=49, FN=82, TP=4040.

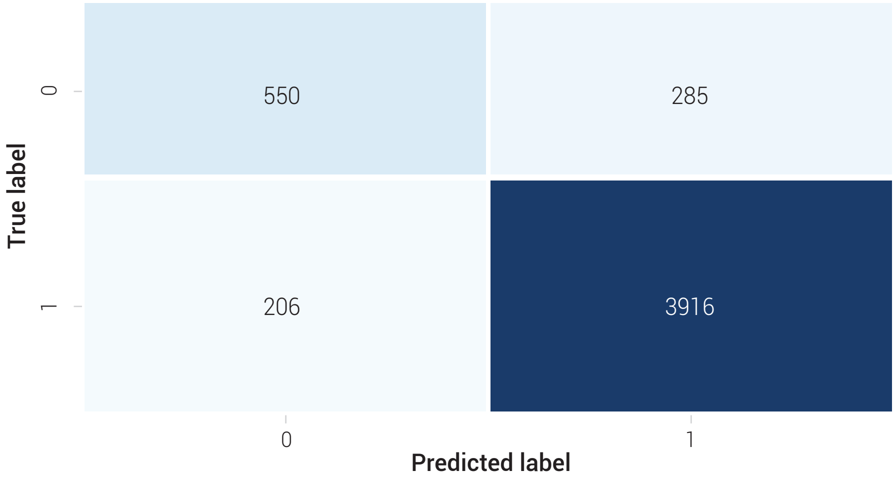

Fig.12 LSTM model confusion matrix
Source: Own work

Fig.12 represents the confusion matrix that was obtained after applying LSTM model on dataset1. This matrix was obtained after applying the LSTM method on the public data set. The TN=550, FP=285, FN=206, TP=3916.

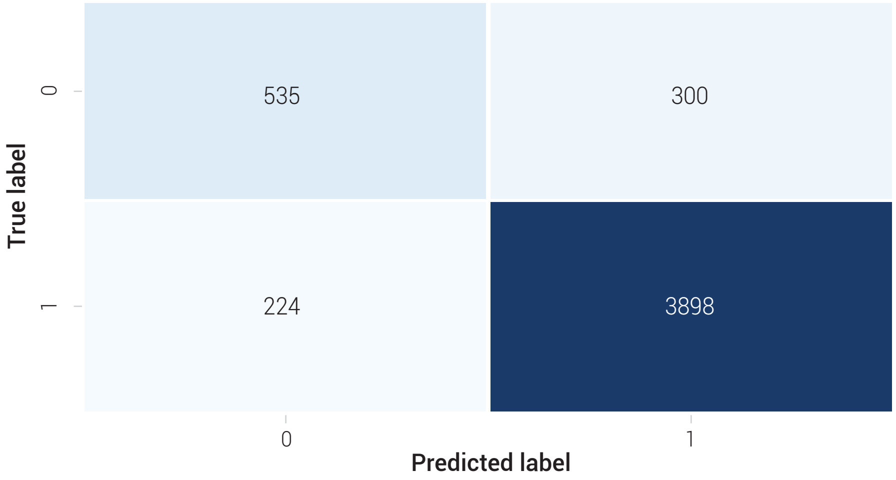

Fig.13 GRU model confusion matrix
Source: Own work

Fig.13 represents the confusion matrix that was obtained after applying the RNN model on dataset1. This matrix was obtained after applying the GRU method on the public data set. The TN=535, FP=300, FN=224, TP=3898.

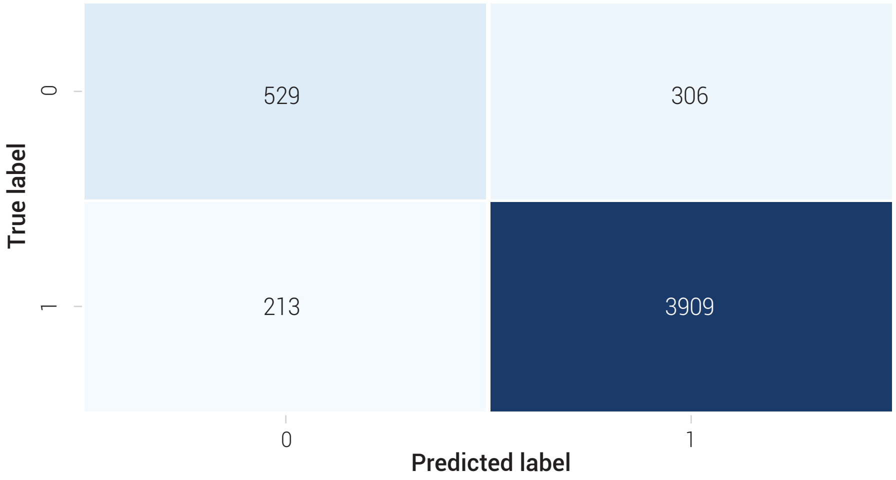

Fig.14 BI- RNN model confusion matrix
Source: Own work

Fig.14 represents the confusion matrix that was obtained after applying the BI-RNN model on dataset1. This matrix was obtained after applying the BI-RNN method on the public data set. The TN=366, FP=49, FN=82, TP=4040.

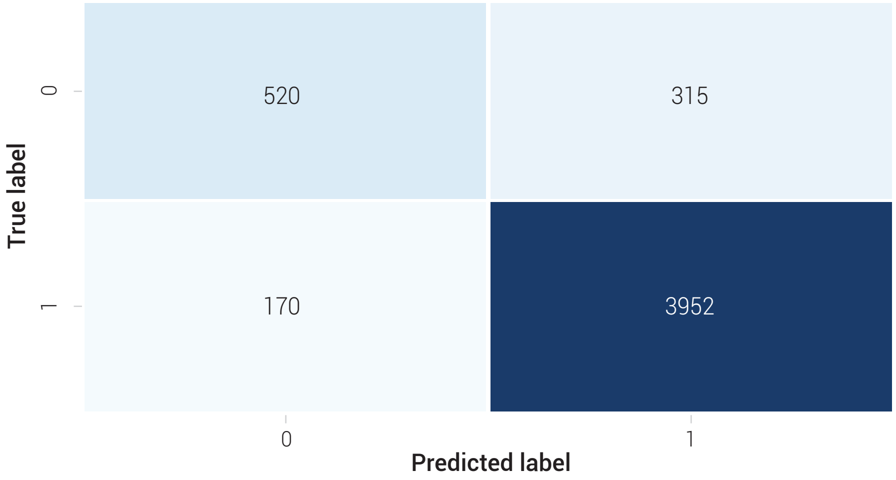

Fig.15 BI-LSTM model confusion matrix

Fig.15 represents the confusion matrix that was obtained after applying the BI-LSTM model on dataset1. This matrix was obtained after applying the BI-LSTM method on the public data set. The TN=520, FP=315, FN=170, TP=3952.

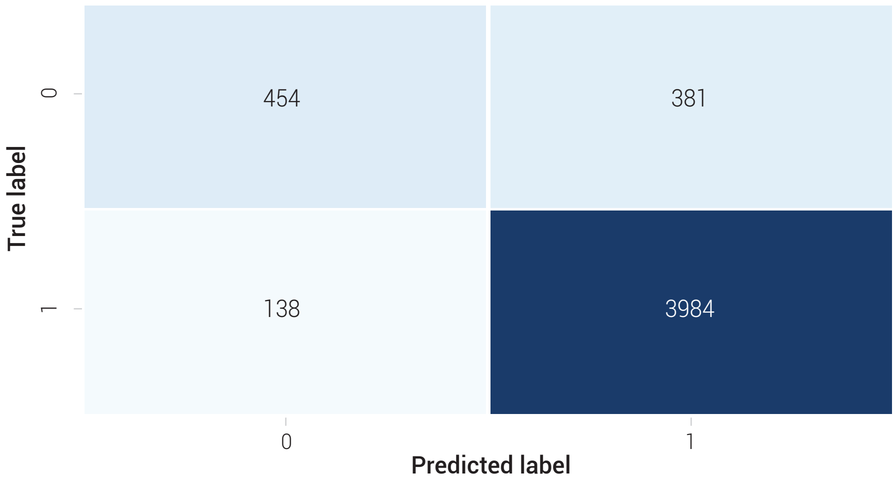

Fig. 16 BI-GRU model confusion matrix
Source: Own work

Fig.16 represents the confusion matrix that was obtained after applying the BI-GRU model on dataset1. This matrix was obtained after applying the BI-GRU method on the public data set. The TN=454, FP=381, FN=138, TP=3984.

Table 4. Comparative results of various deep learning models over Twitter dataset.

Deep Learning Models	Precision	Recall	F1-Score	Accuracy
RNN	0.89	0.89	0.89	89.48
LSTM	0.89	0.90	0.89	89.9
GRU	0.90	0.90	0.90	90.1
BI-RNN	0.88	0.89	0.88	88.76
BI-LSTM	0.90	0.90	0.90	89.73
BI-GRU	**0.90**	**0.90**	**0.90**	**90.43**

Source: Own work

Table 4 represents the results obtained after applying various deep-learning techniques like RNN, BI-RNN, LSTM, BI-LSTM, GRU, and BI-GRU over the Twitter data set. After applying these techniques, accuracy, precision, recall, F1-score were calculated using the confusion matrix obtained. From Table 4 above, the BIGRU method has acquired the greatest accuracy.

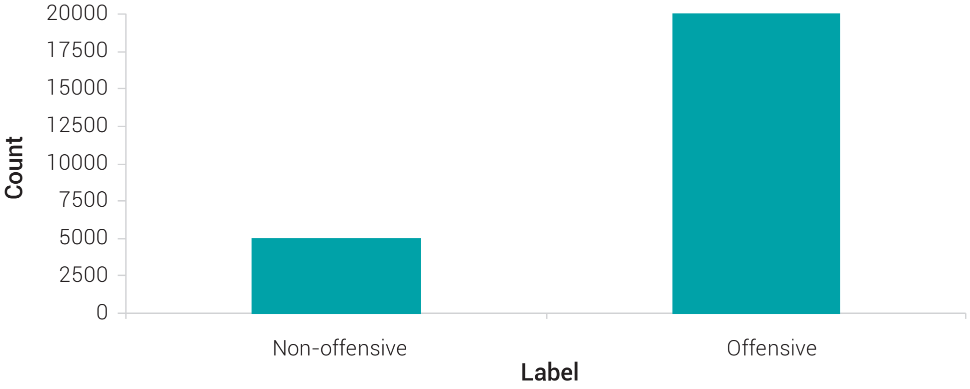

Fig.17 Count of Offensive and non-offensive messages
Source: Own work

Fig.17 represents the Count of offensive and non-offensive comments that are obtained from dataset2. Therefore, from the diagram we can depict that there were over 5000 Non-offensive comments and 20,000 offensive comments from the Twitter dataset.

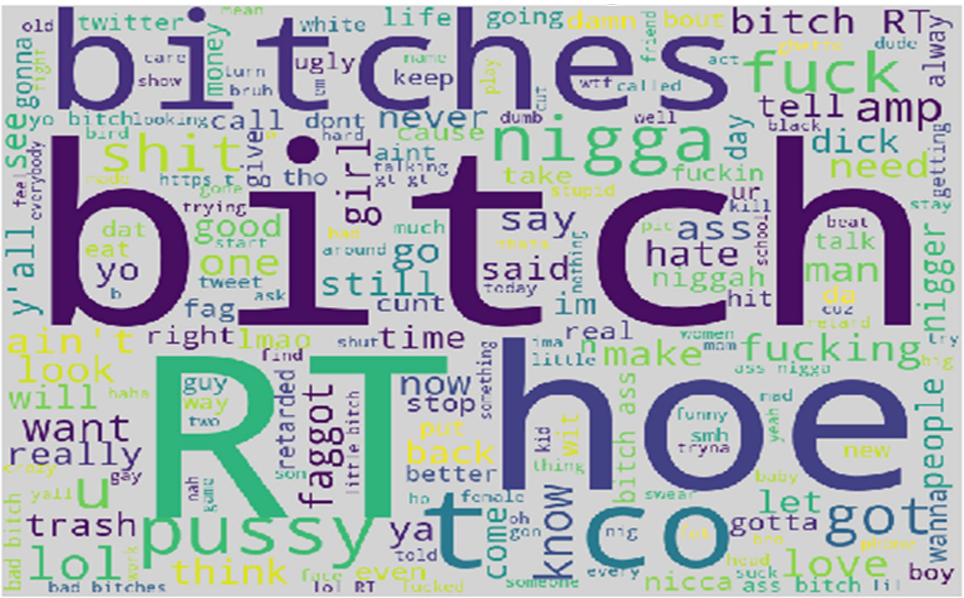

Fig. 18 Offensive messages
Source: Own work

The Fig.18 represents the word cloud of offensive messages that are obtained from the public dataset. Some of the offensive messages are amp, ass, bitch etc. The primary aim of the project was to find out the offensive and non-offensive comments from the data set

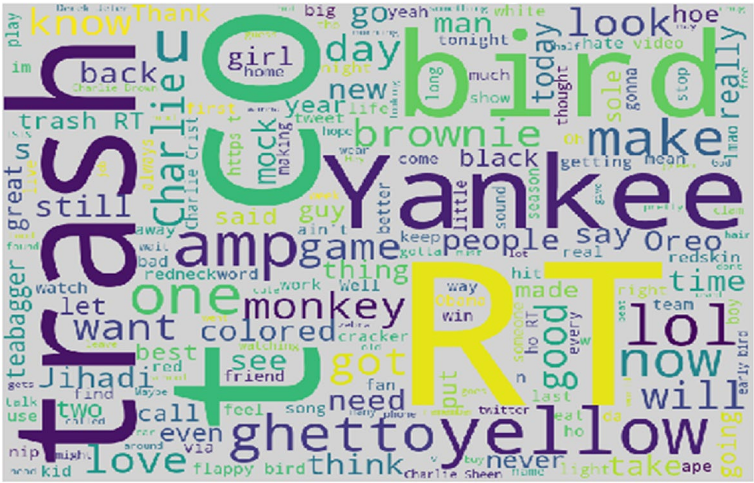

Fig. 19 Non- Offensive messages
Source: Own work

The Fig. 19 represents the Non-offensive messages. Some of the non-offensive messages are trash, bird, love etc. The primary aim of the project was to find out the offensive and non-offensive comments from the data set.

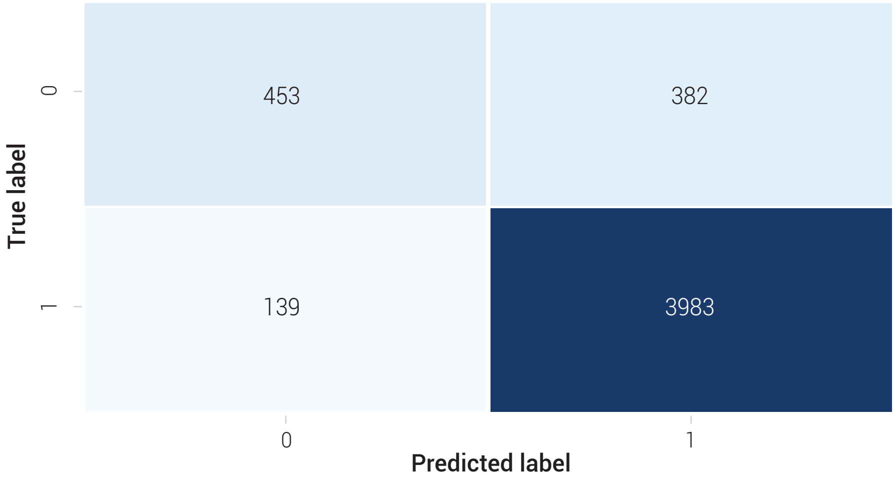

Fig. 20 RNN model confusion matrix
Source: Own work

Fig. 20 represents the confusion matrix that was obtained after applying the RNN model on dataset2. This matrix was obtained after applying the RNN method on the Twitter data set. The TN=453, FP=382, FN=139, TP=3983.

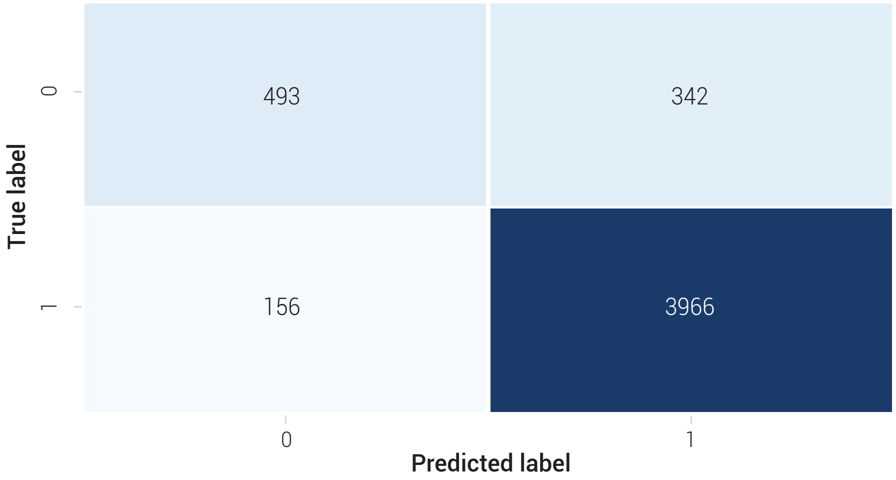

Fig.21 LSTM model confusion matrix
Source: Own work

Fig.21 represents the confusion matrix that was obtained after applying the LSTM model on dataset1. This matrix was obtained after applying the LSTM method on the Twitter data set. The TN=493, FP=342, FN=156, TP=3966.

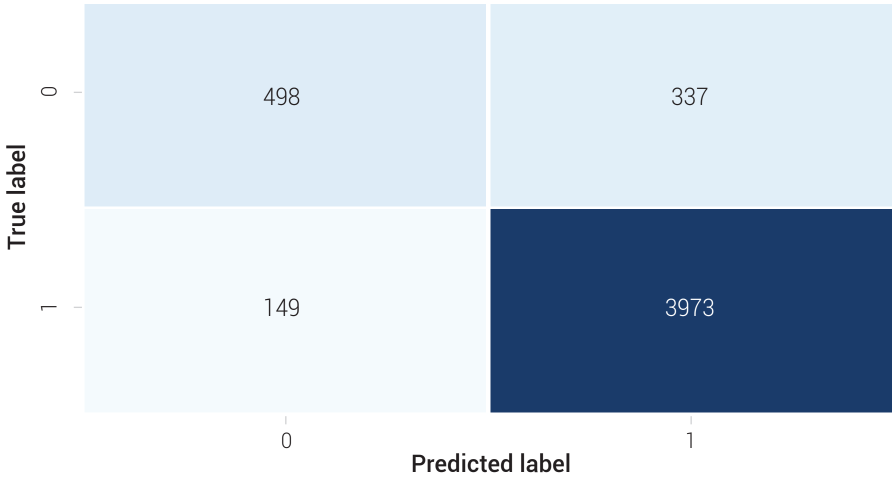

Fig.22 GRU model confusion matrix
Source: Own work

Fig.22 represents the confusion matrix that was obtained after applying the GRU model on dataset2. This matrix was obtained after applying the GRU method on the Twitter data set. The TN=498, FP=337, FN=149, TP=3973.

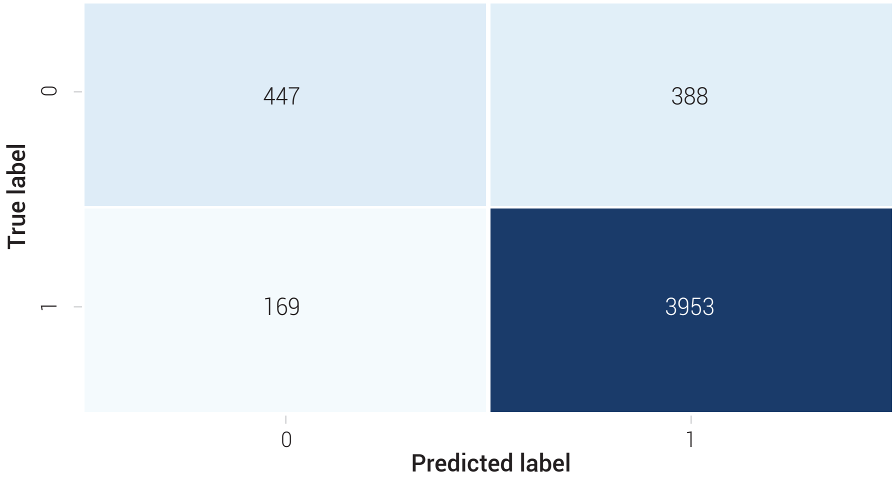

Fig.23 BI-RNN model confusion matrix
Source: Own work

Fig.23 represents the confusion matrix that was obtained after applying the BI-RNN model on dataset2. This matrix was obtained after applying the BI-RNN method on the Twitter data set. The TN=447, FP=388, FN=169, TP=3953.

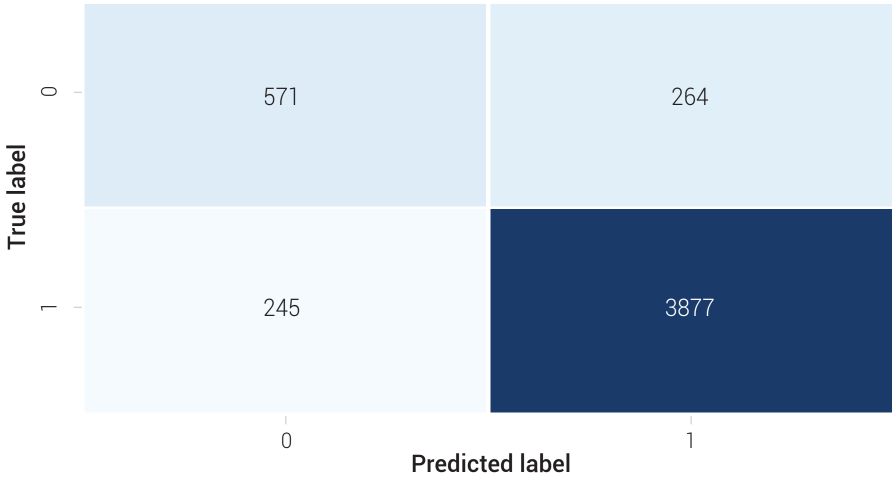

Fig.24 BI-LSTM model confusion matrix
Source: Own work

Fig. 24 represents the confusion matrix that was obtained after applying the BI-LSTM model on dataset2. This matrix was obtained after applying the BI-LSTM method on the Twitter data set. The TN=571, FP=264, FN=245, TP=3977.

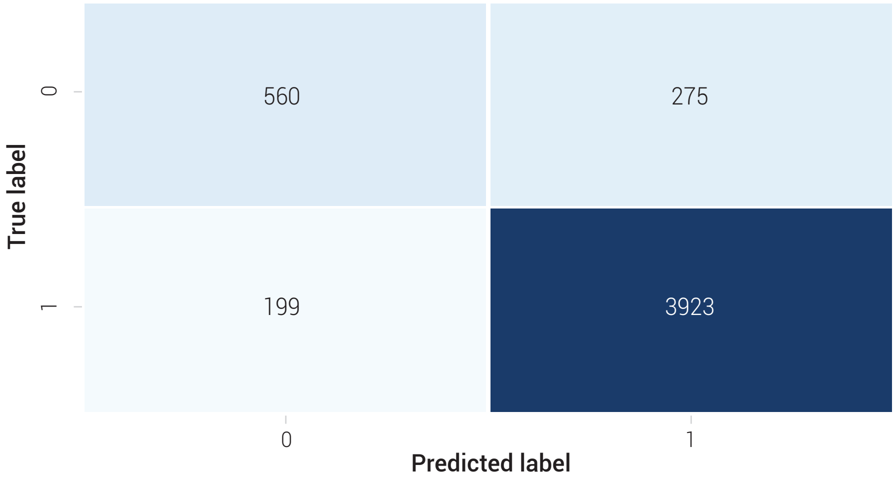

Fig.25 BI-GRU model confusion matrix
Source: Own work

Fig.25 represents the confusion matrix that was obtained after applying the BI-GRU model on dataset2. This matrix was obtained after applying the BI-GRU method on the Twitter data set. The TN=560, FP=275, FN=199, TP=3923.

5. CONCLUSION AND FUTURE WORK

In this paper, we have investigated and compared the various Deep Learning models to detect cyberbullying in social media. The experimental results show that the performance of the proposed methodologies is reasonably good. After applying various deep learning models on the public dataset and Twitter dataset, we have observed an increase in accuracy for the BIGRU model. In future, this work may be extended to the detection of cyberbullying in video, audio, and images datasets.

References

[1] P.K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, “Cyberbullying: Its nature and impact in secondary school pupils,” *Journal of child psychology and psychiatry,* vol. 49*,* no. 4, pp. 376-385. doi: 10.1111/j.1469-7610.2007.01846.x

[2] P. Sayanta, and S. Sriparna, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems,* 2020, pp. 1-8. doi: https://doi.org/10.1007/s00530-020-00710-4

[3] X. Jun-Ming, J. Kwang-Sung, Z. Xiaojin, and A. Bellmore, "Learning from bullying traces in social media," *In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 656-666. 2012. doi: https://aclanthology.org/N12-1084

[4] V. S. Subrahmanian, and K. Srijan, "Predicting human behavior: The next frontiers," *Science,* vol. 355, no. 6324, pp. 489-489. doi: 10.1126/science.aam7032

[5] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily in the digital world: A LiveJournal case study," *IEEE Internet Computing,* vol. 14, no. 2, pp. 15-23. [Online]. Available: Homophily in the Digital World: A LiveJournal Case Study (smu.edu.sg)

[6] M.A. Al-Garadi, D. V. Kasturi, and R. Sri Devi, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior*, vol. 63, pp. 433-443. doi: https://doi.org/10.1016/j.chb.2016.05.051

[7] P. Lawrence, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, "Using social media to predict the future: a systematic literature review," *arXiv preprint* arXiv:1706.06134. doi: https://doi.org/10.48550/arXiv.1706.06134

[8] S. Balakrishna, M. Thirumaran, and V. Kumar Solanki, "A Framework for IoT Sensor Data Acquisition and Analysis," *EAI Endorsed Transactions on Internet of Things*, EAI, vol. 4, no. 16, pp. 1-13. doi: http://dx.doi.org/10.4108/eai.21-12-2018.159410

[9] S. Balakrishna and M. Thirumaran, "Programming Paradigms for IoT Applications: An Exploratory Study", In: Solanki, V. (Ed.), Díaz, V. (Ed.), Davim, J. (Ed.) *Handbook of IoT and Big Data*. Boca Raton: CRC Press, Taylor & Francis Group, Print. February 2019. doi: https://dx.doi.org/10.1201/9780429053290-2

[10] S. Balakrishna, M. Thirumaran, R. Padmanaban, and V. Kumar Solanki, "An Efficient Incremental Clustering-based Improved K-Medoids for IoT Multivariate Data Cluster Analysis," *Peer-to-Peer Networking and Applications*, Springer, vol.13, no.4, pp. 1152-1175. doi: https://dx.doi.org/10.1007/s12083-019-00852-x

[11] S. Balakrishna and M. Thirumaran "Semantic Interoperability in IoT and Big Data for Health-care: A Collaborative Approach," In: Balas V., Solanki V., Kumar R., Khari M. (eds) *A Handbook of Data Science Approaches for Biomedical Engineering*, Elsevier. January 2020. doi: https://dx.doi.org/10.1016/B978-0-12-818318-2.00007-6

[12] D. J. Hemanth, and J. Anitha, "Brain signal based human emotion analysis by circular back propagation and Deep Kohonen Neural Networks," *Computers & Electrical Engineering*, vol. 68, pp. 170-180. doi: https://doi.org/10.1016/j.compeleceng.2018.04.006

[13] Giap, Cu Nguyen, Le Hoang Son, and F. Chiclana, "Dynamic structural neural network," *Journal of Intelligent & Fuzzy Systems,* vol. 34, no. 4, pp. 2479-2490. doi: https://doi.org/10.3233/jifs-171947

[14] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for mode-lling sentences," *arXiv preprint* arXiv:1404.2188. (2014). doi: 10.3115/v1/P14-1062

[15] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *In Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 30, no. 1. 2016. doi: https://doi.org/10.48550/arXiv.1507.04808

[16] A. Rakhlin, "Convolutional Neural Networks for Sentence Classification," GitHub. 2016. https://github.com/alexander-rakhlin/CNN-for-Sentence-Classification-in-Keras

[17] R. Johnson, and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint* arXiv:1412.1058. 2014.doi: https://doi.org/10.48550/arXiv.1412.1058

[18] I. Sutskever, J. Martens, and G.E. Hinton, "Generating text with recurrent neural networks," In ICML. 2011. doi: https://dl.acm.org/doi/10.5555/3104482.3104610

[19] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," *In IJCAI-99 workshop on machine learning for information filtering*, vol. 1, no. 1, pp. 61-67. 1999. doi: maxent.dvi (kamalnigam.com)

[20] L. Pengfei, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," arXiv preprint arXiv:1605.05101. doi: https://doi.org/10.48550/arXiv.1605.05101

[21] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint* arXiv:1308.0850. doi: https://doi.org/10.48550/arXiv.1308.0850

[22] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint* arXiv:1409.1259. doi: https://doi.org/10.48550/arXiv.1409.1259

[23] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, and H. Shakeel, “Deep learning for multi-class identification from domestic violence online posts,” *IEEE Access 7*, pp. 46210-46224. doi: 10.1109/ACCESS.2019.2908827

[24] J. Risch, and R. Krestel, “Aggression identification using deep learning and data augmentation,” *In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018),* pp. 150-158. 2018. doi: https://researchr.org/publication/RischK18-0

[25] D. Nguyen, Kamela Ali Al Mannai, Shafiq Joty, H. Sajjad, M. Imran, and Prasenjit Mitra, “Robust classification of crisis-related data on social networks using convolutional neural networks,” *In Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1. 2017. doi: https://ntunlpsg.github.io/publication/2017_6/

[26] S. Subramani, Hua Wang, Huy Quan Vu, and Gang Li, “Domestic violence crisis identification from Facebook posts based on deep learning,” *IEEE access,* vol. 6, pp.54075-54085. doi: 10.1109/ACCESS.2018.2871446

[27] S. Salawu, Yulan He, and J. Lumsden, “Approaches to automated detection of cyberbullying: A survey,” *IEEE Transactions on Affective Computing,* vol. 11, no. 1 (2017): 3-24. doi: Transaction / Regular Paper Title (aston.ac.uk)

[28] R. Zhao, A. Zhou, and Kezhi Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” In *Proceedings of the 17th international conference on distributed computing and networking*, pp. 1-6. 2016. doi: https://research.aston.ac.uk/en/publications/approaches-to-automated-detection-of-cyberbullying-a-survey/fingerprints/

[29] L. Cheng, Kai Shu, Siqi Wu, Yasin N. Silva, D. L. Hall, and Huan Liu, “Unsupervised cyberbullying detection via time-informed gaussian mixture model,” *arXiv preprint arXiv:2008.02642*. doi: arXiv:2008.02642v1

[30] K. Kumari, Jyoti Prakash Singh, Yogesh Kumar Dwivedi, and Nripendra Pratap Rana. “Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach,” *Soft Computing,* vol. 24, no. 15, pp. 11059-11070. doi: https://doi.org/10.1007/s00500-019-04550-x

[31] M. A. Al-Garadi, M. Rashid Hussain, N. Khan, G. Murtaza, H. Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, H. A. Khattak, and A. Gani, “Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges,” *IEEE Access,* vol. 7, pp. 70701-70718. doi: 10.1109/ACCESS.2019.2918354

[32] A. Kumar, and N. Sachdeva, “Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network,” *Multimedia Systems*, pp.1-10. doi: https://doi.org/10.1007/s00530-020-00747-5

[33] P. Sayanta, and Sriparna Saha, “CyberBERT: BERT for cyberbullying identification,” *Multimedia Systems,* pp. 1-8. doi: https://doi.org/10.1007/s00530-020-00710-4

[34] K. Kumari, and Jyoti Prakash Singh, “Identification of cyberbullying on multi-modal social media posts using genetic algorithm,” *Transactions on Emerging Telecommunications Technologies,* vol. 32, no. 2, pp. e3907. doi: 10.1002/ett.3907

[35] A. Muneer, and Suliman Mohamed Fati, “A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter,” *Future Internet,* vol. 12, no. 11, pp. 187. doi: https://doi.org/10.3390/fi12110187

[36] B.A. Talpur, and D. O’Sullivan, “Cyberbullying severity detection: A machine learning approach,” *PloS one,* vol. 15, no. 10. e0240924. doi: https://doi.org/10.1371/journal.pone.0240924

[37] C. Iwendi, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. “Cyberbullying detection solutions based on deep learning architectures,” *Multimedia Systems*, pp. 1-14. doi: https://doi.org/10.1007/s00530-020-00701-5

[38] N. Lu, Guohua Wu, Zhen Zhang, Yitao Zheng, Yizhi Ren, and Kim-Kwang Raymond Choo, “Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts,” *Concurrency and Computation: Practice and Experience,* vol. 32, no. 23, pp. e5627. doi: 10.1002/cpe.5627

[39] P. Sayanta, and Sriparna Saha, “CyberBERT: BERT for cyberbullying identification,” *Multimedia Systems*, pp. 1-8. doi: https://doi.org/10.1007/s00530-020-00710-4

[40] C. Van Hee, J. Gilles, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. Pauw, W. Daelemans, and V. Hoste, “Automatic detection of cyberbullying in social media text,” *PloS one,* vol. 13, no. 10, pp. e0203794. doi: 10.1371/journal.pone.0203794

[41] D. Van Bruwaene, Qianjia Huang, and D. Inkpen, “A multi-platform dataset for detecting cyberbullying in social media,” *Language Resources and Evaluation,* vol. 54, no. 4, pp. 851-874. doi: https://doi.org/10.1007/s10579-020-09488-3

[42] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, “A “deeper” look at detecting cyberbullying in social networks,” In *2018 international joint conference on neural networks (IJCNN)*, pp. 1-8. IEEE. doi: 10.1109/IJCNN.2018.8489211

[43] A. Kumar, and N. Sachdeva, “Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis,” *Multimedia Tools and Applications,* vol. 78, no. 17, pp. 23973-24010. doi: https://doi.org/10.1007/s11042-019-7234-z

[44] M. Dadvar, and Kai Eckert, “Cyberbullying detection in social networks using deep learning based models; a reproducibility study,” *arXiv preprint arXiv:1812.08046*. doi: https://doi.org/10.48550/arXiv.1812.08046