

# Integrated Approach to Detect Cyberbullying Text: Mobile Device Forensics Data

G. Maria Jones<sup>1,\*</sup>, S. Godfrey Winster<sup>2</sup> and P. Valarmathie<sup>3</sup>

<sup>1</sup>Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, 602105, India

<sup>2</sup>Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Chengalpattu, 603203, India

<sup>3</sup>Department of Information Technology, Saveetha Engineering College, Chennai, 602105, India

\*Corresponding Author: G. Maria Jones. Email: joneofarc26@gmail.com

Received: 14 April 2021; Accepted: 23 May 2021

**Abstract:** Mobile devices and social networks provide communication opportunities among the young generation, which increases vulnerability and cybercrimes activities. A recent survey reports that cyberbullying and cyberstalking constitute a developing issue among youngsters. This paper focuses on cyberbullying detection in mobile phone text by retrieving with the help of an oxygen forensics tool-kit. We describe the data collection using forensics technique and a corpus of suspicious activities like cyberbullying annotation from mobile phones and carry out a sequence of binary classification experiments to determine cyberbullying detection. We use forensics techniques, Machine Learning (ML), and Deep Learning (DL) algorithms to exploit suspicious patterns to help the forensics investigation where every evidence contributes to the case. Experiments on a real-time dataset reveal better results for the detection of cyberbullying content. The Random Forest in ML approach produces 87% of accuracy without SMOTE technique, whereas the value of F1Score produces a good result with SMOTE technique. The LSTM has 92% of validation accuracy in the DL algorithm compared with Dense and BiLSTM algorithms.

**Keywords:** Mobile forensics; cyberbullying; machine learning; investigation model; suspicious pattern

## 1 Introduction

Bullying and Stalking are not novel phenomena to the world. The study states [1] that traditional bullying is limited to place, time, and predictable, whereas cyberbullying tends to happen at any time and place. With the ubiquity of the Internet and social media like blogs, social network sites like Twitter, Facebook and Instant Messaging like Whatsapp, Instagram, Telegram and many more applications make communication with anyone irrespective of place and time. There are two sides to social media: positive sides where people can share useful information and establish social relationships. The second phase of social media is the negative approach, where an increased risk for children with threatening messages, cyberbullying and cyberstalking, etc. The use of advanced techniques to commit cybercrimes is



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

challenging for investigation and evidence collection. So, the use of forensics techniques helps to reveal the offender. The primary usage of digital forensics is to reconstruct and retrieve the digital data from electronic devices, which are utilized for legal proceedings by identifying, analyzing, and capturing the data. The existence of digital information in electronic and digital data represents each and everyone's activity while working, living, and playing in a virtual environment, which creates electronic trace in our daily lives. The evolution of cybercrime in today's world is at an unprecedented phase. In recent years, social networking usage by individuals and companies is drastically rising. This is because there is a rapid growth in smart devices, internet facilities, storage space, etc. The Internet has speed and the ability to transfer the data mainly used for communication purposes and opens the door for criminals to indulge in crimes [2]. In traditional times, criminals usually leave traces of a crime by fingerprints or physical evidence, requiring a short period to investigate. Since technology is increasing rapidly, cybercrimes are also rising exponentially. Most cybercrimes target information about individuals, governments, or corporations.

The data on the computer or network may be modified, merged, or deleted. Digital forensics investigators experience difficulty in gathering the evidence since the criminals use the false identity. Some of the crimes carried out by criminals are hacking, spoofing, phishing, etc. The ultimate goal is to find the veracity, where the evidence has been hidden and has not been discovered, increasing future attacks. Digital Evidence from the sources like social networking services (Whatsapp, WeChat, Line, Instagram, etc.) includes voice calls, SMS, MMS, Audio, Video to show the data breach. The investigator should answer the six essential questions during the investigation: Who, How, What, Why, When, and Where [3]. The information extracted from the compromised device will help to identify the criminals and for the legal proceedings. Mobile Forensics deals with the seizure, acquisition, analysis, and reporting with tools like Encase, Autopsy, Access data, FTK, Oxygen forensics, OSForensics, etc., that can be used to reveal the evidence. Cell phones and Smartphones come under the mobile phone category, which are portable devices. They are vital for day-to-day activities, so they are vulnerable to criminal activity or part of the crime scene. Many smart devices contain user-sensitive information, including their phone call logs, SMS, MMS, electronic mails, photographs, videos, memos, passwords, Web History, and credit/debit card numbers. These device holders use smartphones for communication, exchange photos, connect to social networks, write blogs, record audio, video, etc. Due to technology and transmission, the data rate is at its peak [4]. It allows most individuals to transfer digital data (e.g., digital video, digital images, etc.). Hence, the mobile computing and communication technologies development gives opportunities for criminals and investigators alike.

Cyberbullying and cyberstalking is the dark phase of human nature on a technical side, especially in social media. So, detection becomes a key area for cyberbullying and cyberstalking research. In this work, we propose a framework for cyberbullying detection from mobile text messages using forensics techniques to retrieve the content even if it is deleted. This work aims to help the forensics investigation department to analyze the behavior patterns of victims and offenders. We also present the SMOTE (Synthetic Minority Oversampling Technique) to solve the imbalance problem. Based on the extracted features from messages, we developed ML, DL models for cyberbullying detection. We applied SMOTE technique for ML algorithms and word embedding technique for Dense, LSTM, and BiLSTM models. The features are applied to Logistic Regression, Decision Tree, Random Forest, and XGBoost algorithms.

The paper is organized as follows: Section 2 describes forensics related works for cyberbullying and cyberstalking with sentimental analysis. Section 3 describes the architecture and implementation of integrated method of forensics with ML and DL models. Section 4 consists of Algorithms used for implementation. Section 5 provides the analysis of experimental outcomes compared to other algorithms with and without SMOTE technique. Section 6 provides the result and discussion. Finally, Section 7 provides the conclusion of the study.

## 2 Related Works

In this section, we provide a literature survey and an overview of cyberbullying, sentimental analysis for text, and text forensics analysis. We briefly summarize the forensics based model for cyberbullying and cyberstalking in 2.1 and the sentimental analysis based model in 2.2.

### 2.1 Forensics in Cyberbullying and Cyberstalking

Many works related to cyberbullying and cyberstalking detection rely on both machine learning and deep learning models. The study based on Behavioural Evidence Analysis (BEA) on cyberstalking cases is conducted by Noora et al. [5]. The authors used forensics techniques on 20 cyberstalking cases. They concluded that BEA helped to focus on an investigation that enables better understanding and victim, offender behavior based on digital evidence. The authors [6] have used crowdsourcing techniques to annotate post and hashtags from seven social media platforms to generate cyberbullying data sets. They used Support Vector Machine, XGBoost, and CNN models to perform the experiment. Ingo et al. [7] presented a framework called Anti Cyberstalking Text-based system (ACTS) for detecting text-based cyberstalking. The framework is designed as a prevention mechanism to analyze, detect, identify and block communication. The framework added a forensics technique for collecting evidence. Michael et al. [8] categorized the text and evaluated it using rule-based decision formula and machine learning approach. The authors also used forensics text for deep learning analysis, which help to identify the criminals.

The authors presented a hybrid ontology technology to collect the forensics data from social networks and intended to implement it with advanced operations as future work [9]. The work describes the methodology for retrieving information from Microsoft Skype to identify the end-user devices of a VoIP call by analysing the CODECs exchanged by the clients during the SIP (Session Initiation Protocol) handshaking phase [10]. The author used 7 machine learning algorithms to trace file system, identify how these files can be manipulated and compared with performance measure indicating that neural networks and random forest showed the highest accuracy among these 7 algorithms [11]. This article presents Structural Feature Extraction Methods (SFEM) to detect malicious content in documents by means of three experimental analysis of machine learning algorithms and proposed in future to work on the detection of malicious content in Excel and PowerPoint [12]. The author presented majorclust algorithm to detect suspicious activities in logs which assists forensics examiner to inspect the log files and achieved 70.59%, 82.21% and 83.14% of sensitivity, specificity, and accuracy respectively [13].

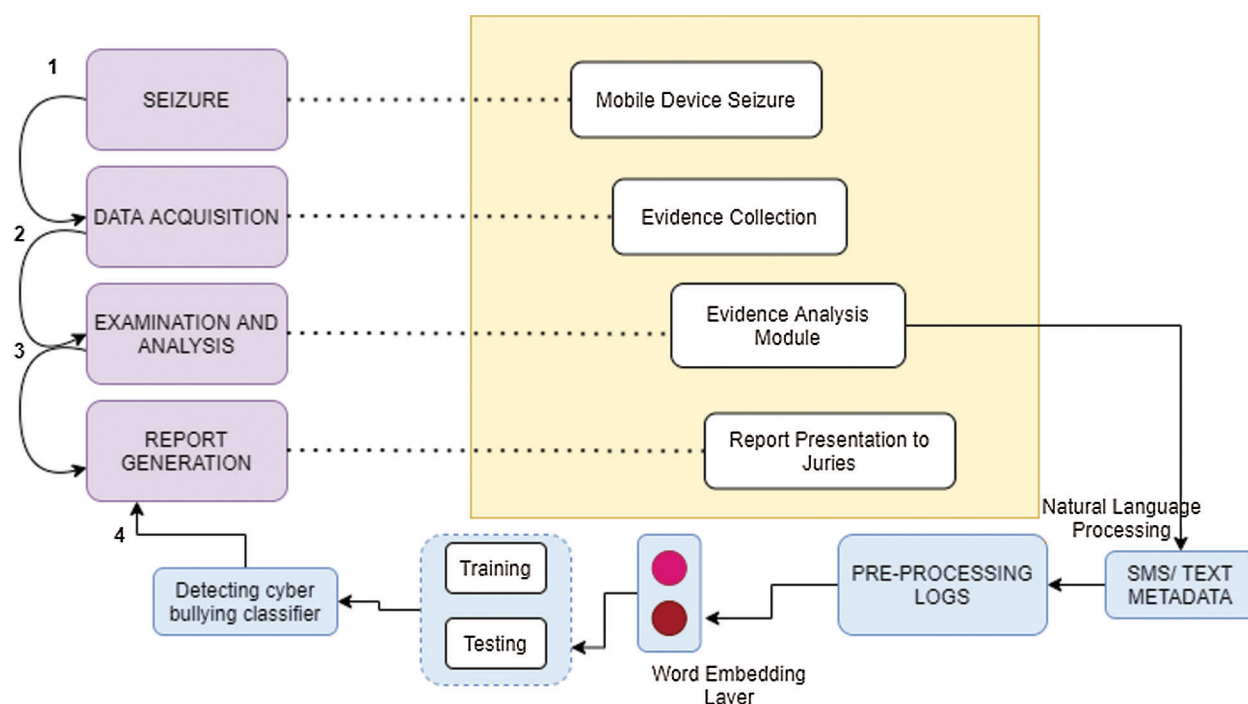
### 2.2 Sentimental Analysis

The authors [14] focused on sentimental analysis to identify the intensity of textual information. The study aimed to classify the social content based on high extreme, moderate, low extreme, and neutral with classification algorithms. Based on the result, Linear SVM performed 82% of accuracy and 88% accuracy in lexicon validation. Bandeh et al. [15] proposed a framework for cyberbullying to generate the feature from Twitter content and used four machine learning algorithms. Finally, the authors have compared the proposed and baseline algorithm of machine learning summarized as a proposed result produced good outcome. Vijayaragavan et al. [16] proposed a new classification model for online product reviews with 1811 instances with two classes. To extract the features, sentimental analysis is used, and finally, a fuzzy-based approach is used to determine product purchases. Sergio et al. [17] used clustering techniques for the publicly available Enron email dataset to analyze the text, which is helpful for digital investigation. Kashfia et al. [18] described how to detect people's emotions and sentiments from their Twitter posts. Their experimental analysis detected six types of emotions. Junseok et al. [19] proposed a new method of weighing and feature selection for Twitter data using sentimental analysis. In this method, the researchers used Naïve Bayes algorithm to estimate the weight, and Multinomial NB was also used to remove the words. The final result produced a good accuracy compared to the existing method.

Gang et al. [20] presented Attention-based Bi-directional Long Term Short Memory with a Convolution layer (AC-BiLSTM) to extract the phrase from word embedding. The final result indicated that AC-BiLSTM performed with good accuracy as compared with other algorithms. Tao et al. [21] revealed the spatial patterns of tweet messages and used the Latent Dirichlet Allocation model to classify the geo-tweet. Duyu et al. [22] introduced the encoding method and sentiment level data simultaneously into the word embedding model and applied a hybrid model to capture context and sentiment data which performed best in all three ways. Lei et al. [23] proposed Sentidiff algorithm to identify the relationship between information in Twitter messages. The data set is demonstrated by a hybrid approach classifier called sentiment classifier and sentiment reversal prediction. The algorithm achieved between 5.09% and 8.38% of PR-AUC. Guixian et al. [24] proposed a method to improve the word representation vector which is an integrated approach of sentiment analysis and TF-IDF. The word vector is given as input to BiSLTM and the study is compared with RNN, CNN, LSTM, and Naïve Bayes. The experimental result showed that the proposed method effectively high accuracy on comments.

### 3 The Integrated Model of Forensics with ML and DL Models

The block diagram of the proposed framework is represented in Fig. 1. The proposed methodology involves three main processes: evidence collection, analyzing the text messages, and performance measure. In the first stage, we collected the evidence from the mobile device in a forensically sound manner. The oxygen forensics software was used to collect the digital evidence from a mobile device (Samsung A50). There are three methods for data acquisition. They are Logical acquisition, Physical acquisition, and manual acquisition. There are about 944 text messages collected from a mobile device through logical acquisition. The text messages are classified as cyberbullying and non-cyber bullying. The dataset contains 168 bully content and 776 content of the non-bully text. We classified annotated cyberbully content into two labels like cyberbullying and non-cyber bullying, to perform the cyberbullying dataset experimental study.



**Figure 1:** Proposed method for detecting cyberbullying text

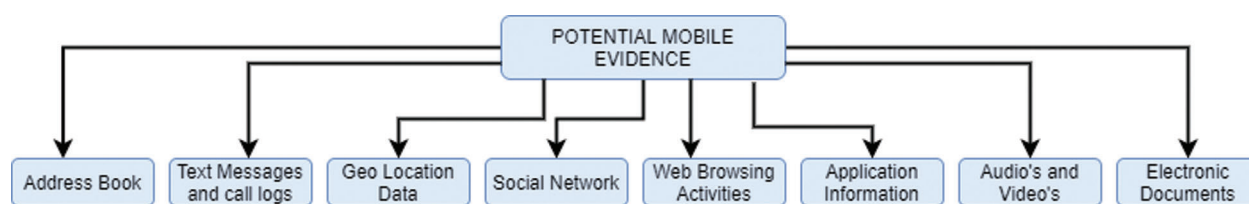
The second stage is analyzing the text message which was retrieved from a mobile device. During analyzing phase, the text is analyzed and categorized as binary classification. The retrieved source of evidence from Samsung A50 is represented in Fig. 3. Finally, the dataset comprises of training and testing in CSV (Comma Separated Value) format. The dataset is classified into 70:30 ratio where each content is text messages. Tab. 1 represents the sample conversation. The final segment is to analyze the text messages with respect to cyberbullying and the ML and DL models are used for training and testing the corpora. To understand the behavior of cyberbullies in text messages, we ran the dataset to understand how ML and DL models identify the bully conversations. This makes us to understand that it would be useful for digital forensics investigators to identify the pattern of offenders.

**Table 1:** Example for chat based cyberbullying

Line	Message	Bully/Not Bully	Type
1	Me and my friends going to party	Not	Neutral
2	F*ck don't you have any boyfriend.	Bully	insult
3	Just kill yourself, I hate	Bully	Curse
4	I will message you later	Not	Neutral
5	U wanna be killed	Bully	Threat

### 3.1 Dataset Collection

Mobile phone users have sensitive information features and capabilities like Personal information storage or management, messaging, audio, video, web browsing, and many more features. These features vary based on the device, developers, and the modification is updating in each version and application installed by users. The following Fig. 2 represents the potential evidence that resides in the mobile phones:



**Figure 2:** Potential Evidence from mobile devices


### 3.2 Steps Involved in Data Preparation

#### 3.2.1 Pre-Processing

Data pre-processing is an essential step to prepare the raw data to analyze the text data. Pre-processing aims to facilitate the training and testing process in machine learning algorithms where the model learns from the data for better results. Some of the steps involved in data pre-processing is discussed below:

#### 3.2.2 Tokenization

Tokenization is the method of breaking down the text into a small entity. During the process, unwanted elements like punctuation are eliminated. Each token is helpful to identify and reveal the pattern of the text document.

<input checked="" type="checkbox"/> <b>Sources</b>	1
<input checked="" type="checkbox"/>  Files	439
<input checked="" type="checkbox"/> <b>Tags</b>	13
<input checked="" type="checkbox"/>  Alcohol	3
<input checked="" type="checkbox"/>  Chat	76
<input checked="" type="checkbox"/>  Child abuse	13
<input checked="" type="checkbox"/>  Currency	5
<input checked="" type="checkbox"/>  Document	333
<input checked="" type="checkbox"/>  Drugs	5
<input checked="" type="checkbox"/>  Extremism	8
<input checked="" type="checkbox"/>  Gambling	4
<input checked="" type="checkbox"/>  Graphic violence	13
<input checked="" type="checkbox"/>  ID / Credit card	14
<input checked="" type="checkbox"/>  Pornography	2
<input checked="" type="checkbox"/>  Vehicles	10
<input checked="" type="checkbox"/>  Weapon	1

**Figure 3:** Evidence retrieved using Oxygen Forensics

The token can divide the document into paragraphs, paragraphs into sentences, and phrases or sentences into words represented as individual words and sentences. The example is tabulated in [Tab. 2](#).

**Table 2:** Example for Tokenization

S.No	Sentence Without Tokenization	Sentence With Tokenization
1	Online Chatting can be tricky	‘Online’, ‘Chatting’, ‘can’, ‘be’, ‘tricky’
2	Harassing or threatening someone	‘Harassing’, ‘or’, ‘threatening’, ‘someone’
3	Pretending to be someone	‘Pretending’, ‘to’, ‘be’, ‘someone’

### 3.2.3 Stemming and Lemmatization

Stemming and Lemmatization is the process of generating the root word from inflected words. The difference between stem and lemma is stem words are not an actual word whereas, a lemma word is an actual language word. Stemming follows an algorithm with steps to perform on the words, which makes it faster. The example for stemming and lemmatization is given in [Tab. 3](#).

**Table 3:** Example for Stemming and Lemmatization

Words	Stem	Lemma
Studies	Studi	Study
Dancing	Danci	Dance
Beautiful	Beauti	Beauty
Corpora	Corpora	Corpus

### 3.2.4 Removing Stopwords, HTML Tags

Stop words are the most commonly used words in every document like “is”, “was”, “the”, “a” and so on. The stop words need to be removed to perform the task as they do not provide any meaning to the sentences as mentioned in Tab. 4. Before training the machine learning and deep learning models, the stopwords are often removed from the dataset, increasing the time efficiency and the performance measures.

**Table 4:** Example for Stopwords and HTML tags

With Stop Words	Without Stop Words
Online Chatting can be tricky	Online, Chatting, tricky
Harassing or threatening someone	Harassing, threatening, someone
Pretending to be someone	Pretending, someone

## 3.3 Feature Extraction

### 3.3.1 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is used to identify the frequency of words present in the document. The term frequency used to measure the frequency of the words present, and it can be formulated as below,

$$Tf(t, d) = \text{No. of words 't' in the document 'd' / total no. of words 't' in 'd'}$$

where,

T represents words present in the document

D represents the document.

The next term is document frequency which is similar to term frequency. The difference among them is that the term frequency analyses the words ‘t’ in document ‘d’ whereas document frequency counts the number of occurrence words ‘t’ in the document. In order words, it can define as count the number of documents in which a word is present. Next, IDF is termed as Inverse Document Frequency which is used to measure the term's information. The following formula is used to calculate the small corpus of data,

$$Idf(t) = S(\text{Document set}) / DF$$

In the case of the huge corpus, the log can be used to calculate as represented below,

$$Idf(t) = \log(S / DF + 1)$$

Finally, by taking multiplicative of TF and IDF, we get TF-IDF as below,

$$TF - IDF = tf(t, d) * \log(S / (DF + 1))$$

### 3.3.2 BOW (Bag of Words)

Bag of words (BOW) is the method of changing the text into fixed-length vectors by the occurrence of terms present in the text document. It is one way for feature extraction from the text for implementing machine learning and deep learning algorithms.

For example, consider two documents;

D1: After the exam, let's have a party at my house

D2: Today's exam is challenging. Let's have some break and go to a party

After eliminating the stop words, the matrix can be formed using unique words from all the documents as given below in [Tab. 5](#).

**Table 5:** BOW representation

	After	Exam	party	tough	Break	House	Today's
D1	1	1	1	0	1	0	0
D2	0	1	1	1	1	0	1

## 4 Considered Algorithms

### 4.1 Machine Learning Algorithms

A comparative analysis was done on the cyberbullying dataset using four classifier algorithms: Logistic Regression, Decision Tree, XGBoost and Random Forest. Since the dataset is imposed with imbalance, SMOTE's resampling technique (Synthetic Minority Oversampling Technique) as mentioned in Algorithm 1 which is used to have a balanced dataset. In this technique, the synthetic samples are created to a less labeled class, and also it helps avoid the overfitting problem. The SMOTE technique is applied to an original dataset with respect to four classification algorithms. The evaluation metrics are compared with and without SMOTE technique to analyze the better output and following algorithms are used for binary classification.

Logistic Regression

Decision Tree

Random Forest

XGBoost

---

#### Algorithm 1: SMOTE Algorithm

#SMOTE Algorithm

Input: Minority data M, Majority data N, Nearest Neighbor n

Output: M

```

1:for S = 1 to M do
2:Compute n Nearest Neighbor for s
3:While M!=N do
4: Choose one nearest neighbor for k
5: Computer vector and multiply the random number
6: Synthetic Data=k + vector
7: End while
8: End for
9: Return M

```

---

### 4.2 Deep Learning Algorithms

The three various deep learning algorithms such as Dense, LSTM and BiLSTM are used to analyze cyberbullying dataset. In dense architecture, the first layer is an embedding layer set as 16 and the input

length is fifty. In order to avoid the overfitting problem, a pooling layer is used. The dense network is defined with the activation function called “relu” with a dropout layer to avoid the overfitting problem, and a final output layer is fixed with a sigmoid function.

The variant of RNN (Recurrent Neural Network) is defined as Long Term Short Memory (LSTM). The main aim of designing the structure is to avoid long-term dependency and vanishing gradient problems due to which the network stops learning. In RNN, the repeating mode of  $\tan_n$  takes place in simple structure, whereas in LSTM, the repeating way takes place in the various structures, and the representation is presented in Fig. 6. In this work, the input series is given as  $x=[x_1, x_2, x_3, \dots, x_n]$  where  $x$  is a word vector and a hidden vector is represented by  $hd_t$  at a period of  $t$ . To learn the cyberbullying classification, we used the LSTM model. Initially, the model learns the hidden vector which is given in the input series and generates the target output based on historical data. The key component of LSTM is  $cd_t$  termed as candidate state or cell state and the model can take decision whether the cell can be modified or added to the memory by using sigmoid gate which is in three forms: input gate ( $ip_t$ ), forget gate ( $frt_t$ ) and output gate ( $op_t$ ).

In forget gate, the sigmoid is set to be executed initially. The LSTM cell chooses how significant the past state in the cell  $C_{t-1}$  is and, at that point, chooses which new data are saved in the  $cd_t$  cell state. This segment has two parts: the  $ip_t$  (input gate) will decide what information to be updated and next, the  $\tan_n$  layer makes a vector of  $c_{t-1}$  (candidate layer) included to the state. Finally, the decision can be taken to remove the data. Now, the next stage is to update the old state  $c_{t-1}$  to a new cell state. Then, the old state can be multiplied by  $frt_t$  and add it by  $it \cdot c_t$ .

This generates the new candidate cell. The final stage is to compute the LSTM model result that can be carried out using the sigmoid and  $\tan_n$  layers. The final result is based on the information that resides in cell state and it also a sigmoid layer used to filter which decides which part of a cell will affect the final output result. Finally, the cell state value applied to the  $\tan_n$  filter and multiplied by the sigmoid layer's output and the formula is described in Eq. (1):

$$\begin{aligned}
 frt_t &= \sigma(w_{fk}[h_{t-1}, x_t] + b_f) \\
 ip_t &= \sigma(w_{ik}[h_{t-1}, t] + b_i) \\
 \dot{c}_t &= \tan_n(w_{ck}[h_{t-1}, x_t] + b_c) \\
 cd_t &= f_t * c_{t-1} + i_t * \dot{c}_t \\
 op_t &= \sigma(w_{ok}[h_{t-1}, x_t] + b_o) \\
 hd_t &= o_t * \tan_n h(c_t)
 \end{aligned} \tag{1}$$

where  $\sigma$  is an activation function which ranges from 0 to 1, i.e., that information can be removed completely, partially removed or completely stored,  $\dot{c}_t$  is abbreviated as candidate hidden state which is computed based on present input values and past hidden state, it is defined as input gate which defines the amount of newly computed state for present input values,  $h_{t-1}$  is termed as recurrent of past and present hidden layer,  $W$  is weight,  $c$  is the internal memory cell and  $hd_t$  is the output state.

Unlike LSTM, BiLSTM works in back-propagation, which means the propagation takes places in both forward and backward directions as represented in Fig. 7. It learns the pattern from before and after with two independent LSTM. It sums up the data from two directions of a sentence and merges the sentimental data. More precisely, At each period of step ‘ $t$ ’, the forward LSTM computes the hidden state ‘ $f$ ’  $h_t$  based on the past state  $f h_{t-1}$  with vector  $x_t$ . Meanwhile, the backward propagation computes the  $xh_t$  based on the  $xh_{t-1}$  with the same vector  $x_t$ . Finally, both the result combined together as the final hidden state. Due to this, the computation time is increased as compared to LSTM. The final result of the BiLSTM model is as follows in (2),

$$h_t = [fh_t, xh_t] \quad (2)$$

## 5 Analysis

### Performance Measure

The standard measure to evaluate the system performances are Precision, Recall, Accuracy, and F1\*score. A confusion matrix is used to measure the correctness and accuracy of the model. Primarily it is used for classification problems. There are four terms associated with the confusion matrix, as mentioned below and in [Tab. 10](#).

TPc (True Positive of cyberbullies): The actual class and predicted class samples are true (1).

TNc (True Negative of Cyberbullies): The samples of cyberbullying actual and predicted class are false (0).

FPc (False Positive of cyberbullies): The sample of the actual class is false (0) and the predicted class is true (1).

FNc (False Negative of cyberbullies): The sample of the actual class is true (1) and the predicted class is false (0).

Accuracy is termed as the ratio of correct prediction to the all prediction made in the classification as represented in (3). Precision is the process of measuring the true samples to the positive samples as represented in (4). The recall is the measure of correctly classified samples to the total number of class and it can be calculated by (5). In F1Score, the harmonic mean is used to calculate as given in (6).

$$Accuracy(A) = \left\langle \frac{TP_c + TN_c}{TP_c + FP_c + TN_c + FN_c} \right\rangle, k \in S_{cb} \quad (3)$$

$$Precision(P) = \left\langle \frac{TP_c}{TP_c + FP_c} \right\rangle, k \in S_{cb} \quad (4)$$

$$Recall(R) = \left\langle \frac{TP_n}{TP_c + FN_c} \right\rangle, k \in S_{cb} \quad (5)$$

$$F1 * Score = \left\langle \frac{2 * PR}{P + R} \right\rangle, k \in S_{cb} \quad (6)$$

From [Eqs. \(3\)–\(6\)](#), we describe how performance measures are computed for all the classes k in the dataset that belongs to the  $S_{cb}$  set of suspicious in cyberbullying.

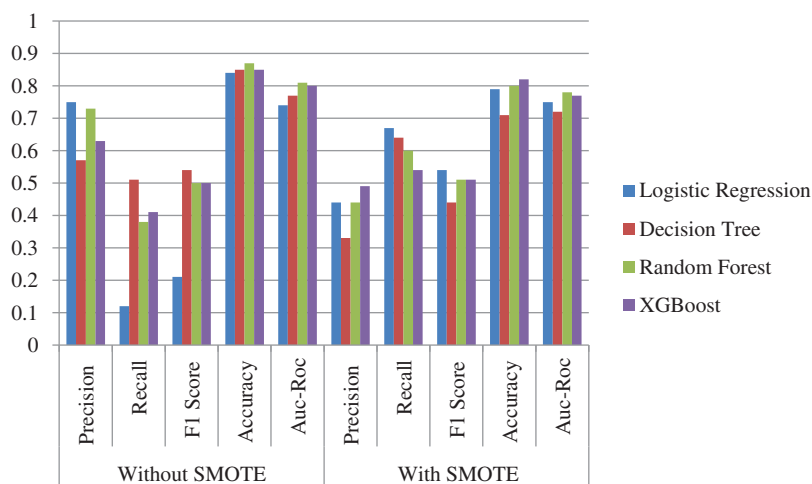
## 6 Result and Discussion

This article comprises three main stages, and the performance measure of ML and DL algorithms is evaluated, which are explained in this session. The first stage is data collection from smartphones using mobile forensics toolkit and the acquisition results vary from the forensics toolkit. We generated the data using oxygen forensics software, performing logical acquisition. The acquired data were analyzed using NLP, and SMOTE techniques were used to perform imbalanced data. The integrated process of Forensics, Machine Learning and Deep Learning process helps to analyze user patterns and produces better results. The second stage is associated with the text processing for better performance. Initially, the four classification algorithms of ML and DL, such as Logistic Regression, Decision Tree, Random Forest, XGBoost, Dense, LSTM and BiLSTM are used. In third stage, the better performance was calculated in terms of Accuracy, Precision, Recall, F1\*score and Auc-Roc. Since the data is imbalanced, the SMOTE technique is also performed and compared with actual dataset.

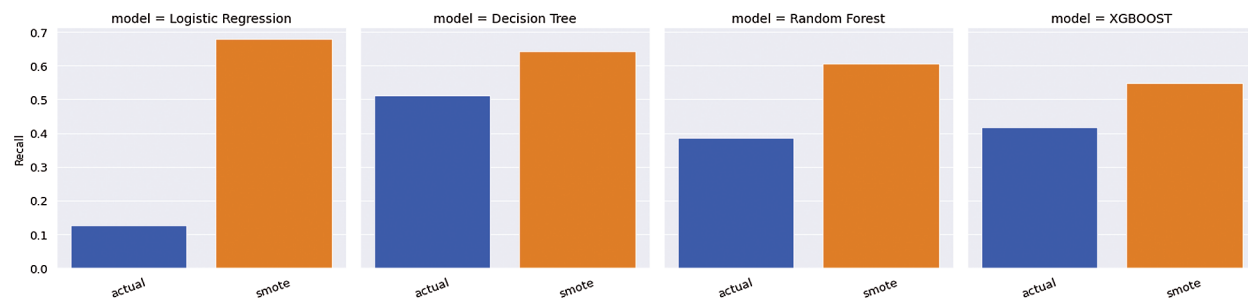
Based on the performance measure depicted in [Tab. 6](#) and [Fig. 4](#), the Random Forest performs the highest accuracy with the rate of 87% followed by XGBoost, Decision Tree and Logistic Regression with 85%, 85% and 84%, respectively, for actual data. In contrast, in SMOTE technique, also XGBoost reached the highest accuracy of 82%, followed by Random Forest, Logistic Regression and Decision Tree with 80%, 79% and 71%, respectively. Comparatively, in terms of accuracy, without smote technique, the algorithm performed well, whereas, in terms of Recall, Smote technique performed well with the rate of 61% for XGBoost, 63% for Random Forest, 63% for Decision Tree and 71% for Logistic Regression as represented in [Fig. 5](#).

**Table 6:** Performance measure of proposed work

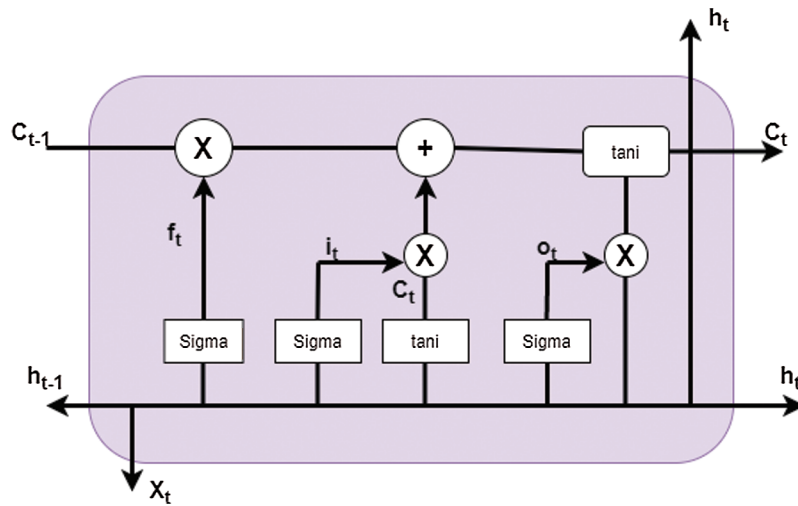
	Algorithm	resample	Accuracy	Precision	Recall	F1-score	AUC-ROC
0	Logistic Regression	actual	0.84	0.750000	0.125000	0.214286	0.746717
1	Logistic Regression	smote	0.79	0.448819	0.678571	0.540284	0.753835
2	Decision Tree	actual	0.85	0.577181	0.511905	0.542587	0.773188
3	Decision Tree	smote	0.71	0.336449	0.642857	0.441718	0.725370
4	Random Forest	actual	0.87	0.738636	0.386905	0.507812	0.816665
5	Random Forest	smote	0.80	0.447368	0.607143	0.515152	0.782792
6	XGBOOST	actual	0.85	0.630631	0.416667	0.501792	0.805090
7	XGBOOST	smote	0.82	0.491979	0.547619	0.518310	0.776617



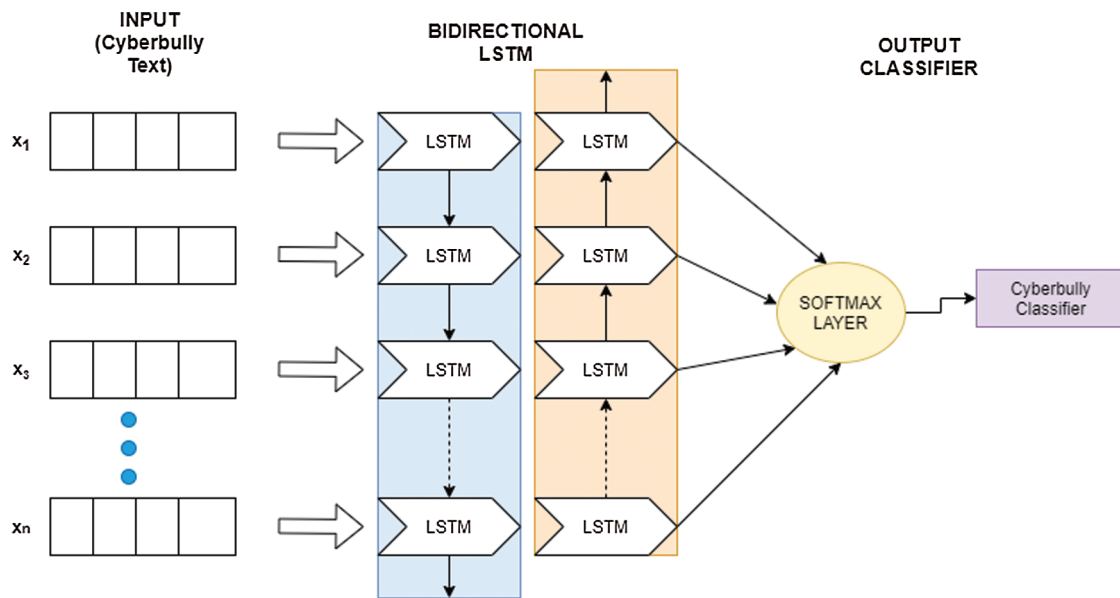
**Figure 4:** Comparison of Actual and SMOTE classification



**Figure 5:** Recall measures between SMOTE and without SMOTE



**Figure 6:** Structure of LSTM



**Figure 7:** Structure of BiLSTM

The final segment of the work deals with deep learning analysis for identifying and classification of cyberbullying data. The data is classified as 70% for training and 30% for testing. Based on the three algorithms evaluation, LSTM performed well where validation accuracy reached 92% and 68%, 89% for Bilstm, presented in [Tabs. 8 and 9](#); [Figs. 9 and 10](#). The Dense algorithm reached about 84% of validation accuracy, as tabulated in [Tab. 7](#) and represented in [Fig. 8](#).

**Table 7:** Performance measure for Dense Network

Training_Loss	Training_Accuracy	Validation_Loss	Validation_Accuracy
0.686263	0.702092	0.676125	0.842809
0.657547	0.859414	0.631415	0.849498

**Table 8:** Performance Measure of LSTM

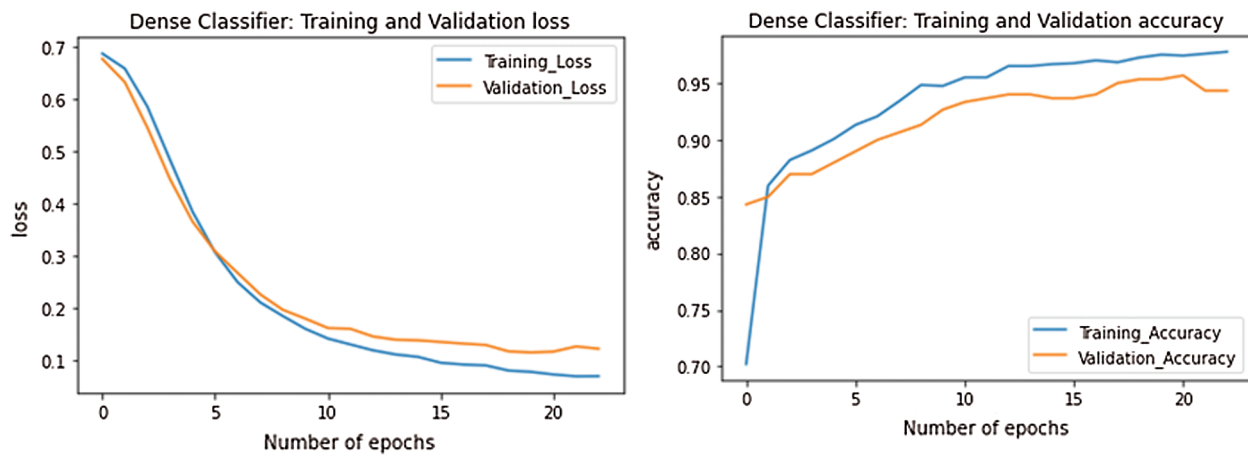
Training_Loss	Training_Accuracy	Validation_Loss	Validation_Accuracy
0.139294	0.952486	0.258179	0.925485
0.136586	0.953992	0.293780	0.921271

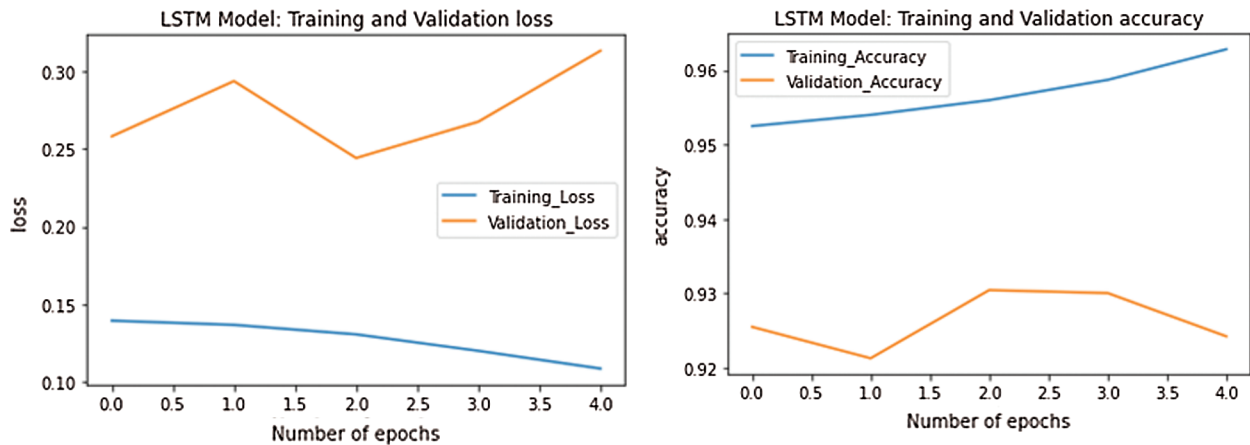
**Table 9:** Performance Measure of biLSTM

Training_Loss	Training_Accuracy	Validation_Loss	Validation_Accuracy
0.672027	0.605992	0.605110	0.680803
0.412528	0.859732	0.312381	0.892642

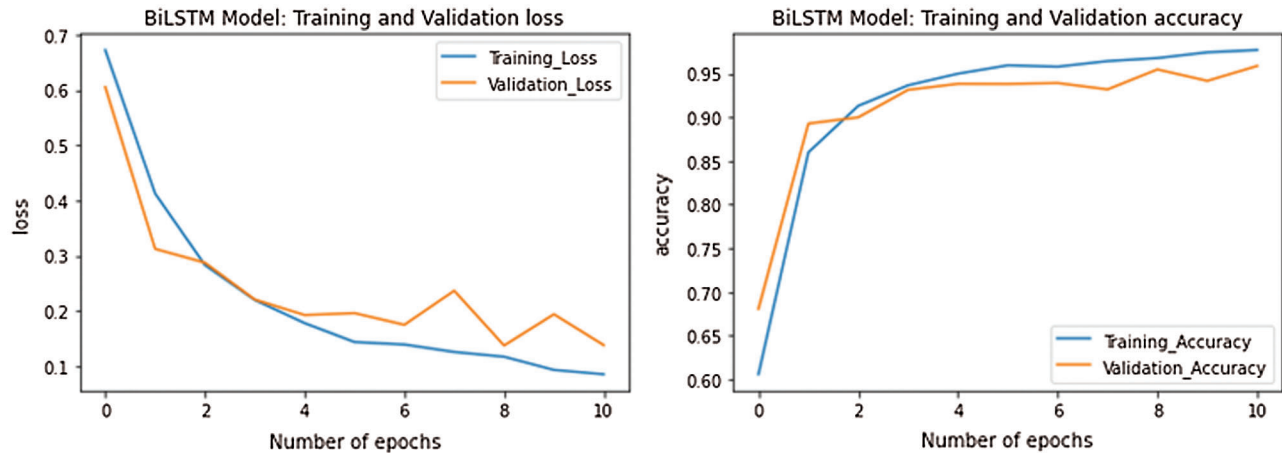
**Table 10:** Confusion matrix for performance metrics

Predicted Value	Actual Value	
	1	0
Positive (1)	True Positive (TPc)	False Positive (FPc)
Negative (0)	False Negative (FNc)	True Negative (TNc)

**Figure 8:** Performance of Dense classifier



**Figure 9:** Performance of LSTM



**Figure 10:** Performance of BiLSTM

The experimental result showed that implementing forensics data to ML and DL models can provide better results when investigating cyberbullying cases. In this scenario, usually, the physical crime scene and evidence is absent. So, in cybercrime activities, the computer, laptops, mobile devices and tablets are considered essential evidence sources. Like in the physical crime scene, the offender leaves the virtual evidence that can be inferred and analyzed using digital forensics methodology. In many cases, the offender uses sufficient skills to hide the traces. In that case, the behavioral patterns can help to distinguish the pattern from other contents.

## 7 Conclusion and Future Work

This study examined the behavioral pattern of cyberbullies in the context of a digital forensics investigation. Text analysis and machine learning approach with forensics data are new techniques towards cybercrime investigation or incident response teams. This integrated approach helps to identify the behavioral patterns of victim and offender and solve many criminal cases. Many times, the Internet and social media usage lead to the involvement in cyberbullying and cyberstalking. In this paper, we developed a framework for detecting and identifying the pattern of cyberbullies. The forensics technique has been used to retrieve text messages from mobile phones and it is pre-processed using NLP

techniques. The four ML model is developed and compared with SMOTE technique. The accuracy of ML reached 87% using Random Forest, whereas using SMOTE, the recall value in XGBoost reached the highest. The three deep learning algorithms are also performed in which LSTM reached the highest validation accuracy compared to Dense and BiLSTM. In this regard, future work aims to develop a new mechanism for automatic detection of harassment, threats, hate, and stalking content of offenders. With the help of ML and DL models, the investigation team can get an accurate pattern of the victim and offender.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] M. Cepin, U. Slana, B. Roggenbuck, V. Edert, M. Kaps, G. Trevisan et al. How is Cyberbullying different from Traditional bullying?. 2016. [Online]. Available: <http://socialna-akademija.si/joiningforces/3-2-how-is-cyberbullying-different-from-traditional-bullying/>.
- [2] G. M. Jones and S. G. Winstler, "Forensics analysis on smart phones using mobile forensics tools," *International Journal of Computational Intelligence Research*, vol. 13, no. 8, pp. 1859–1869, 2017.
- [3] D. Quick and K. R. Choo, "Pervasive social networking forensics: Intelligence and evidence from mobile device extracts," *Journal of Network and Computer Application*, vol. 86, pp. 24–33, 2017.
- [4] G. M. Jones and S. G. Winstler, "Analysis of crime report by data analytics using python," in *challenges and applications of data analytics in social perspectives*, IGI Global, Hershey, PA 17033, USA, pp. 54–79, 2020.
- [5] N. Al, J. Bryce, V. N. L. Franqueira and A. Marrington, "Forensic investigation of cyberstalking cases using behavioural evidence analysis," *Digital Investigation*, vol. 16, pp. S96–S103, 2016.
- [6] D. V. Bruwaene and Q. Huang, "A multi-platform dataset for detecting cyberbullying in social media," *Language Resource Evaluation*, vol. 54, no. 4, pp. 851–874, 2020.
- [7] I. Frommholz, H. M. Martin, P. Zinnar, G. Mitul and S. Emma, "On textual analysis and machine learning for cyberstalking detection," *Datenbank Spektrum*, vol. 16, pp. 127–135, 2016.
- [8] M. Spranger and D. Labudde, "Semantic Tools for Forensics: Approaches in Forensic Text Analysis," in *Proc. IMMM*, pp. 97–100, 2013.
- [9] H. Arshad, A. Jantan, G. Keng and A. Sahar, "A multilayered semantic framework for integrated forensic acquisition on social media," *Digital Investigation*, vol. 29, pp. 147–158, 2019.
- [10] M. Nicoletti and M. Bernaschi, "Forensic analysis of microsoft skype for business," *Digital Investigation*, vol. 29, pp. 159–179, 2019.
- [11] R. M. A. Mohammad and M. Alqahtani, "Journal of information security and applications a comparison of machine learning techniques for file system forensics analysis," *Journal of Information Security and Application*, vol. 46, pp. 53–61, 2019.
- [12] A. Cohen, N. Nissim, L. Rokach and Y. Elovici, "SFEM: Structural feature extraction methodology for the detection of malicious office documents using machine learning methods," *Expert System Application*, vol. 63, pp. 324–343, 2016.
- [13] H. Studiawan, C. Payne and F. Sohel, "Graph clustering and anomaly detection of access control log for forensic purposes," *Digital Investigation*, vol. 21, pp. 76–87, 2017.
- [14] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid and J. Shah, "Telematics and informatics sentiment analysis of extremism in social media from textual information," *Telematics Informatics*, vol. 48, pp. 1–20, 2020.
- [15] B. Ali, T. Id and D. O. Sullivan, "Cyberbullying severity detection: A machine learning approach," *PLOS One*, vol. 15, no. 10, pp. 1–19, 2020.
- [16] P. Vijayaragavan, R. Ponnusamy and M. Aramudhan, "An optimal support vector machine based classification model for sentimental analysis of online product reviews," *Future Generation Computer System*, vol. 111, pp. 234–240, 2020.

- [17] S. Decherchi, S. Tacconi, J. Redi, F. Sangiacomo, A. Leoncini *et al.*, “Text clustering for digital forensics analysis,” In: Herrero Á., Gastaldo P., Zunino R., Corchado E. (Eds). *Computational Intelligence in Security for Information Systems. Advances in Intelligent and Soft Computing*, vol. 63. Heidelberg: Springer, 2009.
- [18] K. Sailunaz and R. Alhajj, “Emotion and sentiment analysis from twitter text,” *Journal of Computational Science*, vol. 36, pp. 101003, 2019.
- [19] J. Song, K. T. Kim, B. Lee, S. Kim and H. Y. Youn, “A novel classification approach based on naïve Bayes for twitter sentiment analysis,” *TIIS*, vol. 11, no. 6, pp. 2996–3011, 2017.
- [20] G. Liu and J. Guo, “Bidirectional LSTM with attention mechanism and convolutional layer for text classification,” *Neurocomputing*, vol. 337, pp. 325–338, 2019.
- [21] B. She and L. Duan, “A systematic spatial and temporal sentiment analysis on geo-tweets,” *IEEE Access*, vol. 8, pp. 8658–8667, 2020.
- [22] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu *et al.*, “Sentiment embeddings with applications to sentiment analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 1–14, 2016.
- [23] L. Wang, J. Niu, S. Member, S. Yu and S. Member, “Sentidiff: Combining textual information and sentiment diffusion patterns for twitter sentiment snalysis,” *IEEE Transaction Knowledge Data Engineering*, vol. 14, no. 8, pp. 1–14, 2019.
- [24] G. Xu, “Sentiment analysis of comment texts based on BiLSTM,” *IEEE Access*, vol. 7, pp. 51522–51532, 2019.