International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

CYBERBULLYING PREVENTION ON SOCIAL NETWORKING SITES USING DEEP LEARNING MODEL

M.MOHAMMED IJAS[1], MUTHU KUMARAN D[2], PRASANTH T[3], Ms.K.PRADEEPA[4], ME

[1]*Computer Science & E.G.S .Pillay Engineering College, Nagapattinam*
[2]*Computer Science & E.G.S .Pillay Engineering College, Nagapattinam*
[3]*Computer Science & E.G.S .Pillay Engineering College, Nagapattinam*
[4]*Computer Science & E.G.S .Pillay Engineering College, Nagapattinam*

Abstract - Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation. The content an individual share online – both their personal content as well as any negative, mean, or hurtful content – creates a kind of permanent public record of their views, activities, and behaviour. To avoides detecting cyberbullying attacks, many existing approaches in the literature incorporate Machine Learning and Natural Language Processing text classification models without considering the sentence semantics. The main goal of this project is to overcome that issue. This project proposed a model LSTM - CNN architecture for detecting cyberbullying attacks and it used word2vec to train the custom of word embeddings.This model is used to classify tweets or comments as bullying or non-bullying based on the toxicity score. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. A convolutional neural network (CNN) is a type of artificial neural network and it has a convolutional layer to extract information by a larger piece of text and by using this model LSTM- CNN achieve a higher accuracy in analysis, classification and detecting the cyberbullying attacks on posts and comments.

***Key Words*:** cyberbullies, bullier, social media, harassment, harm full- content, social media crime, bullying, deep learning.

1 .INTRODUCTION

1.1.OVERVIEW:

Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation. Some cyberbullying crosses the line into unlawful or criminal behaviour.

The most common places where cyberbullying occurs are:

- Social Media, such as Facebook, Instagram, Snapchat, and Tik Tok
- Text messaging and messaging apps on mobile or tablet devices
- Instant messaging, direct messaging, and online chatting over the internet
- Online forums, chat rooms, and message boards, such as Reddit
- Email
- Online gaming communities

1.1.1. Different Kinds of Cyberbullying

There are many ways that someone can fall victim to or experience cyberbullying when using technology and the internet. Some common methods of cyberbullying are:

Harassment – When someone is being harassed online, they are being subjected to a string of abusive messages or efforts to contact them by one person or a group of people. People can be harassed through social media as well as through their mobile phone (texting and calling) and email. Most of the contact the victim will receive will be of a malicious or threatening nature.

Doxing – Doxing is when an individual or group of people distribute another person's personal information such as their home address, cell phone number or place of work onto social media or public forums without that person's permission to do so. Doxing can cause the victim to feel extremely anxious and it can affect their mental health.

Cyberstalking – Similar to harassment, cyberstalking involves the perpetrator making persistent efforts to gain contact with the victim, however this differs from harassment – more commonly than not, people will cyberstalk another person due to deep feelings towards that person, whether they are positive or negative. Someone who is cyberstalking is more likely to escalate their stalking into the offline world.

Revenge porn – Revenge porn, is when sexually explicit or compromising images of a person have been distributed onto social media or shared on revenge porn specific websites without their permission to do so. Normally, images of this nature are posted by an ex-partner, who does it with the purpose of causing humiliation and damage to their reputation.

Swatting – Swatting is when someone calls emergency responders with claims of dangerous events taking place at an address. People swat others with the intention of causing panic and fear when armed response units arrive at their home or place of work. Swatting is more prevalent within the online gaming community.

Corporate attacks – In the corporate world, attacks can be used to send masses of information to a website in order to take the website down and make it non-functional. Corporate attacks can affect public confidence, damaging businesses reputations and in some instances, force them to collapse.

Account hacking – Cyberbullies can hack into a victim's social media accounts and post abusive or damaging messages. This can be particularly damaging for brands and public figures.

False profiles – Fake social media accounts can be setup with the intention of damaging a person or brand's reputation. This can easily be done by obtaining publicly available images of the victim and making the account appear as authentic as possible.

Slut shaming – Slut shaming is when someone is called out and labelled as a "slut" for something that they have done previously or even just how they dress. This kind of cyberbullying often occurs when someone has been sexting another person and their images or conversations become public. It is seen more commonly within young people and teenagers but anyone can fall victim to being slut shamed.

1.1.2. INTRODUCTION:

Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation. Some cyberbullying crosses the line into unlawful or criminal behaviour.

The most common places where cyberbullying occurs are:

- Social Media, such as Facebook, Instagram, Snapchat, and Tik Tok
- Text messaging and messaging apps on mobile or tablet devices
- Instant messaging, direct messaging, and online chatting over the internet
- Online forums, chat rooms, and message boards, such as Reddit
- Email
- Online gaming communities

Social media networks such as Facebook, Twitter, Flickr, and Instagram have become the preferred online platforms for interaction and socialization among people of all ages. While these platforms enable people to communicate and interact in previously unthinkable ways, they have also led to malevolent activities such as cyber-bullying. Cyberbullying is a type of psychological abuse with a significant impact on society. Cyber-bullying events have been increasing mostly among young people spending most of their time navigating between different social media platforms. Particularly, social media networks such as Twitter and Facebook are prone to CB because of their popularity and the anonymity that the Interne provides to abusers. In India, for example, 14 percent of all harassment occurs on Facebook and Twitter, with 37 percent of these incidents involving youngsters. Moreover, cyberbullying might lead to serious mental issues and adverse mental health effects. Most suicides are due to the anxiety, depression, stress, and social and emotional difficulties from cyber-bullying events. This motivates the need for an approach to identify cyberbullying in social media messages (e.g., posts, tweets, and comments).In this project, we mainly focus on the problem of cyberbullying detection on the Twitter platform. As cyberbullying is becoming a prevalent problem in Twitter, the detection of cyberbullying events from tweets and provisioning preventive measures are the primary tasks in battling cyberbullying threats. Therefore, there is a greater need to increase the research on social networks-based CB in order to get greaterinsights and aid in the development of effective tools and approaches to effectively combat cyberbullying problem. Manually monitoring and controlling cyberbullying on Twitter platform is virtually impossible. Furthermore, mining social media messages for cyberbullying detection is quite difficult. For example, Twitter messages are often brief, full of slang, and may include emojis, and gifs, which makes it impossible to deduce individuals' intentions and meanings purely from social media messages. Moreover, bullying can be difficult to detect if the bully uses strategies like sarcasm or passive-aggressiveness to conceal it. Despite the challenges that social media messages bring, cyberbullying detection on social media is an open and active research topic. Cyberbullying detection within the Twitter platform has largely been pursued through tweet classification and to a certain extent with topic modelling approaches. Text classification based on supervised machine learning (ML) models are commonly used for classifying tweets into bullying and non-bullying tweets.Also, it may be suitable only for a pre-determined collection of events, but it cannot successfully handle tweets that change on the fly. Considering these limitations, an efficient tweet classification approach must be developed to bridge the gap between the classifier and the topic model so that the adaptability is significantly proficient.The deep learning model has been successfully used by current researchers to address various issues related to the text domain, including sentiment analysis, question answering, document classification, sentence classification, spam filtering and others. By following them, this project also uses a deep learning algorithm to address the cyber bullying detection issue. BiLSTM is capable of capturing the semantics of the sentence by performing the convolution operations over the tweets.

1.1.3. OBJECTIVE OF THE PROJECT

This project aims to automatically detect cyberbullying from tweets by using deep learning approaches.The aim of this project is to identify the maximum number of cyberbullying related tweets from Twitter as soon as it is posted by users.The objective of our solution is to identify the bullies from raw Twitter data based on the context as well as the contents in which the tweets exist. To warn and block the bully

2. LITERATURE SURVEY

1.Accurate Cyberbullying Detection and Prevention on Social Media

Authors: Andrea Pereraa ,Pumudu Fernando

Year: 2021

Link:

https://www.sciencedirect.com/science/article/pii/S1877050921002507

Objective:

The aim of this project isa system for automatic detection and prevention cyberbullying considering the main characteristics of cyberbullying such as Intention to harm an individual, repeatedly and over time and using abusive curl language or cyberbullying using supervised machine learning.

Methodology:

The usage of digital/social media is increasing day by day with the advancement of technology. People in the twenty-first century are being raised in an internet-enabled world with social media. Communication has been just one button click. Even though there are plenty of opportunities with digital media people tend to misuse it. People spread hatred toward a person in social networking. Cyberbullying affects people in different aspects. It doesn't affect only for health, there are more different aspects which will lead life to a threat. Cyberbullying is a worldwide modern phenomenon which humans cannot avoid hundred percent but can be prevented. Most existing solutions have shown techniques/approaches to detect cyberbullying, but they are not freely available for end-users to use. They haven't considered the evolution of language which makes a big impact on cyberbullying text. This article proposed a TF-IDF (Term Frequency, Inverse Document Frequency) by using TFIDF which can measure the importance of words in a document and Common words such as "is", "am" do not affect the results due to IDF. This article used Support Vector Machines (SVM), A well-known efficient binary classifier to train the model. Logistic regression was used to select the best combination of features. SVM algorithm, training data is used to learn a classification function. It can classify new data not previously seen in one of the two categories. It separates the training data set into two categories using a large hyperplane. Logistic regression is a linear classifier that predicts the probabilities.

Merits:

- Accuracy is high.
- It can classify new data not previously seen.
- Efficiency is high.

Demerits:

- High cost.
- Long and tedius job.

1. Cyberbullying Detection Through Sentiment Analysis

Authors: Jalal Omer Atoum

Year: 2020

Link:https://ieeexplore.ieee.org/document/9458024

Objective:

The aim of this project is a SA model for identifying cyberbullying texts in Twitter social media.

Methodology:

In recent years with the widespread of social media platforms across the globe especially among young people, cyberbullying and aggression have become a serious and annoying problem that communities must deal with. Such platforms provide various ways for bullies to attack and threaten others in their communities. Various techniques and methodologies have been used or proposed to combat cyberbullying through early detection and alerts to discover and/or protect victims from such attacks.This article proposed an approach to detect cyberbullying from Twitter social media platform based on Sentiment Analysis that employed machine learning techniques; namely, Naïve Bayes and Support Vector Machine. The data sets used in this article is a collection of tweets that have been classified into positive, negative, or neutral cyberbullying. Before training and testing such machine learning techniques, the collected set of tweets have gone through several phases of cleaning, annotations, normalization, tokenization, named entity recognition, removing stopped words, stemming and n-gram, and features selection. The results of the conducted experiments have

indicated that SVM classifiers have outperformed NB classifiers in almost all performance measures over all language models. Specifically, SVM classifiers have achieved an average accuracy value of 92.02%, while, the NB classifiers have achieved an average accuracy of 81.1 on the 4-gram language model.

Merits:

- It has better performance measures than NB classifiers on such tweets.
- Low cost and low time consuming.

Demerits:

- Suffer from an inability to detect indirect language harassment.
- Accuracy is low.

2. **Cyberbullying Detection on Social Networks Using Machine Learning Approaches**

Authors: MdManowarul Islam; Md Ashraf Uddin; Linta Islam; ArnishaAkter; Selina Sharmin; Uzzal Kumar Acharjee

Year: 2020

Link:https://ieeexplore.ieee.org/document/9411601

Objective:

The aim of this project is to design and develop an effective technique to detect online abusive and bullying messages by merging natural language processing and machine learning.

Methodology:

The use of social media has grown exponentially over time with the growth of the Internet and has become the most influential networking platform in the 21st century. However, the enhancement of social connectivity often creates negative impacts on society that contribute to a couple of bad phenomena such as online abuse, harassment cyberbullying, cybercrime and online trolling. Cyberbullying frequently leads to serious mental and physical distress, particularly for women and children, and even sometimes force them to attempt suicide. Online harassment attracts attention due to its strong negative social impact. Many incidents have recently occurred worldwide due to online harassment, such as sharing private chats, rumours, and sexual remarks. Therefore, the identification of bullying text or message on social media has gained a growing amount of attention among researchers. The purpose of this article is to design and develop an effective technique to detect online abusive and bullying messages by merging natural language processing and machine learning. Two distinct features, namely Bag-of - Words (BoW) and term frequency-inverse text frequency (TFIDF), are used to analyse the accuracy level of four distinct machine learning algorithms.

Merits:

- The words that occur more frequently should be given more importance as they are more useful for classification.
- Classifier is a supervised learning model which provides accurate result because several decision trees are merged to make the outcome.

Demerits:

- Weak-supervision loss.
- Classifier accuracy is low.

3. **A Fairness-Aware Fusion Framework for Multimodal Cyberbullying Detection**

Authors: Jamal Alasadi; RamanathanArunachalam; Pradeep K. Atrey; Vivek K. Singh

Year: 2020

Link: https://ieeexplore.ieee.org/document/9232508

Objective:

The aim of this project is a fairness-aware fusion framework that ensures that both fairness and accuracy remain important considerations when combining data coming from multiple modalities.

Methodology:

Recent reports of bias in multimedia algorithms (e.g., lesser accuracy of face detection for women and persons of colour) have underscored the urgent need to devise approaches which work equally well for different demographic groups. Hence, here posit that ensuring fairness in multimodal cyberbullying detectors (e.g., equal performance irrespective of the gender of the victim) is an important challenge**.** This article describes one of the first attempts at a Bayesian fusion framework that not only optimizes for accuracy but also considers fairness. The framework takes into account the accuracy and the fairness score for each modality to assign them weights. The weights

of each modality and the agreement between them is used to come up with optimal decisions that balance accuracy and fairness. The results of applying the framework to a multimodal (visual + textual) cyberbullying detection problem demonstrate the efficacy of the approach in yielding high levels of both accuracy and bias. The results pave way for a more accurate and fair approach for cyberbullying detection, which would provide equitable opportunities to different groups in improving their quality of life.

Merits:

- It ensuring both accuracy and fairness.
- Efficiency is high.
- Speed is high to detect the offensive words.

Demerits:

- Algorithms that perform differently for different groups.
- Many issue of fairness for not detecting the cyberbullying cases.
- Time consuming process.

4. LSHWE: Improving Similarity-Based Word Embedding with Locality Sensitive Hashing for Cyberbullying Detection

Authors: Zehua Zhao; Min Gao; Fengji Luo; Yi Zhang; QingyuXiong

Year: 2020

Link: https://ieeexplore.ieee.org/document/9207640

Objective:

The aim of this project is a word embedding method called LSHWE to solve this limitation, which is based on an idea that deliberately obfuscated words have a high context similarity with their corresponding bullying words.

Methodology:

This articleproposes a similarity-based word embedding method LSHWE to solve the "deliberately obfuscated words" problem in cyberbullying detection task. LSHWE has two steps. Firstly, for a given corpus, it generates: (a) a co-occurrence matrix C; (b) a rare word list R; (c) a nearest neighbour list NL obtained by locality sensitive hashing; and (d) a nearest neighbour matrix N. Secondly, an LSH-based auto encoder is used to learn the word vectors according to C and N. The proposed embedding method has two characteristics: (1) LSHWE can represent well on rare words. LSHWE is a global similarity-based word embedding method thus the representations of rare words learnt through LSHWE can be as close as possible to their corresponding words' representations; and (2) LSHWE is a highly efficient algorithm. This method uses an approximate nearest neighbour search method to search the top-k nearest neighbours instead of exact nearest neighbour search methods, which can greatly reduce the running time. This article design experiments from three aspects: effectiveness of LSHWE on cyberbullying detection task, algorithm efficiency and parameter sensitivity. Experiment results demonstrate that LSHWE can alleviate the "deliberately obfuscated words" problem and is highly efficient on large-scale datasets.

Merits:

- It can alleviate the "deliberately obfuscated words" problem.
- Highly efficient on large-scale datasets.

Demerits:

- Running time is long.
- High Cost.

5. Artificial Bee Colony–Based Feature Selection Algorithm for Cyberbullying

Authors: EsraSaracEssiz; Murat Oturakci

Year: 2019

Link:https://ieeexplore.ieee.org/document/9433175

Objective:

The aim of this project is toexamine the effects of the ABC-based feature selection algorithm on classification performance for cyberbullying.

Methodology:

ABC-based feature selection system was proposed for the cyberbully detection problem. IG, CHI2 and Relief methods are used as traditional feature selection methods while classifications are obtained by using Weka data mining tool for the experiments. The proposed ABC-based feature selection method is coded in Java NetBeans platform. Experimental evaluation shows that the proposed method is over performed with respect to traditional filter-based feature selection methods. The classification performance has been increased significantly over the baseline result that is shown in

Table 3. The Macro averaged Fmeasure of the data set is increased from 0.659 to 0.8 using ABC-based feature selection method. The proposed method is effective in reducing the number of features so that it is suitable for classification of high dimensional data. Proposed system also reduces the time required to classify test data set sharply without loss of accuracy in classification performance with reduced feature space.

Merits:

- Without loss of accuracy in classification performance.
- Reduces the time required to classify test data set sharply.
- Effective in reducing the number of features so that it is suitable for classification of high dimensional data.

Demerits:

- Costly and practically an unlikely process.
- Large error for learning classification model.

6. Cyberbullying Detection on Twitter using Multiple Textual Features

Authors: Jianwei Zhang; Taiga Otomo; Lin Li; Shinsuke Nakajima

Year: 2019

Link:https://ieeexplore.ieee.org/document/8923186

Objective:

The aim of this project is to focus on the Japanese text on Twitter and construct an optimal model for automatic detection of cyberbullying by extracting multiple textual features and investigating their effects with multiple machine learning models.

Methodology:

This article aims at automatic cyberbullying detection and utilize machine learning methods to realize the purpose. Two most important aspects for classifying cyberbullying based on machine learning are what features to be extracted and what machine learning models to be selected. This article focus on the Twitter text (hereinafter referred to as tweet) and intend to find the features that mostly contribute to cyberbullying detection. This article uses the technique of text mining and analyse a range of textual features including n-gram, Word2Vec, Doc2Vec, tweets' emotion values, and unique characteristics on Twitter. In addition, multiple machine learning models including linear models, tree-based models and deep learning models are investigated with multiple textual features to construct an optimal model. Based on the collected tweets, it evaluates the quality of automatic detection of cyberbullying, and find that the best model with predictive textual features and it achieve the accuracy of over 90%.

Merits:

- Accuracy is high.
- The quality of automatic cyberbullying detection is high.

Demerits:

- Not able to handle large datasets.
- Classification accuracy is low.

7. Cyberbullying Detection on Instagram with Optimal Online Feature Selection

Authors: Mengfan Yao; Charalampos Chelmis; Daphney-Stavroula Zois

Year: 2019

Link:https://ieeexplore.ieee.org/document/8508329

Objective:

The aim of this project is a novel algorithm to drastically reduce the number of features used in classification for cyberbullying detection.

Methodology:

Cyberbullying has emerged as a large-scale societal problem that demands accurate methods for its detection in an effort to mitigate its detrimental consequences. While automated, data-driven techniques for analysing and detecting cyberbullying incidents have been developed, the scalability of existing approaches has largely been ignored. At the same time, the complexities underlying cyberbullying behaviour (e.g., social context and changing language) make the automatic identification of "the best subset of features" to use challenging. To address this gap by formulating cyberbullying detection as a sequential hypothesis testing problem. Based on this formulation, this article proposes a novel algorithm to drastically reduce the number of features used in classification. This article demonstrates the utility, scalability and responsiveness of this article using a real-world dataset

from Instagram, the online social media platform with the highest percentage of users reporting experiencing cyberbullying. This article approach improves recall by a staggering 700%, while at the same time reducing the average number of features by up to 99.82% compared to state-of-the-art supervised cyberbullying detection methods, learning approaches that require weak supervision, and traditional offline feature selection and dimensionality reduction techniques.

Merits:

- Improves recall by a staggering 700%.
- Reducing the average number of features.
- Accuracy is high.

Demerits:

- It cannot be readily used for online classification as it is an offline approach that provides no classification strategy.
- Time consuming process.

8. Cyberbullying Detection using Recursive Neural Network through Offline Repository

Authors: Nidhi Chandra; Sunil Kumar Khatri; Subhranil Som

Year: 2019

Link:https://ieeexplore.ieee.org/document/8748570

Objective:

The aim of this project is to predict the user behaviour based on his posts on social networking sites specifically Twitter.

Methodology:

Internet hides user identity and in a sense, it provides anonymity to user. With the rise in social media among users, societies across the world are now closely connected, sharing their views and ideas in a form of comments, tagging and video sharing. Many of these views are direct views in a form of user opinions or indirect views in a form of pictures and videos, audios and sometimes music (religious or otherwise) where in it is difficult to understand opinion through programmatic means. These contents many of times are misused by unscrupulous elements to brainwash masses against political institutions, creating instability in the society and on many times leading to online radicalization. Such a trend is quite dangerous as it is leading to brain washing of masses without physical presence of such persons. This article is to demonstrate the identification of specific text from the data which is available in various forms – structured and unstructured and coming from various online sources in real time, posted by users worldwide. The online sources referred to in this article are social networking sites, twitter etc. where multiple users collaborate with each other and post contents. To demonstrate the approach, the Information is captured from these sites through API exposed by these vendors which is captured in NoSQL databases and then NLP is applied to break the text which is then applied against text corpus to identify analogous data. From programming perspective data strictures are used to store the data at run-time. Specific WordNet API is being leveraged for their capabilities to find synonyms.

Merits:

- Training the model to perform the auxiliary tasks well will result in good.
- Accuracy and speed to detection is high.

Demerits:

- Low accuracy.
- Time consuming process.

9. Automated Cyberbullying Detection using Clustering Appearance Patterns

Authors: WalisaRomsaiyud; KodchakornnaNakornphanom; P impakaPrasertsilp; PiyapornNurarak; PiromKonglerd

Year: 2017

Link:https://ieeexplore.ieee.org/document/7886127

Objective:

The aim of this project is to enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering.

Methodology:

This article developed an automatic cyberbullying detection system to detect, identify, and classify cyberbullying activities from the large volume of streaming texts from OSN services. Texts are fed into cluster and discriminant analysis stage which is able to identify abusive texts. The abusive texts are then clustered by using K-Mean. Naïve Bayes is used as classification algorithms to build a classifier from the training datasets and build a predictive model. Moreover, it also used

Naïve Bayes to classify the abusive texts into one of the eight pre-defined categories. The categories include activities approach, communicative, desensitization, compliment, isolation, personal information, reframing, and relationship. The proposed approach consists of two main methods. The first method aims to clean and pre-process the datasets by removing non-printable and special characters, reducing the duplicate words and clustering the datasets. The second one concerns classification model to predict the text messages for preventing cyberbullying. The method was executed on Cybercrime Data, which is a manually labelled dataset, for 170,019 posts and Twitter web site for 467 million Twitter posts.

Merits:

- Is able to classify abusive messages from sentences frequency by using statistics score and partition data sources.
- Classification and prediction model is high.

Demerits:

- Accuracy is low.
- Difficult to track the process.

3.PROPOSED SYSTEM

In this paper, we design a model based on the bidirectional BiLSTM to detect cyberbullying in textual form.

- **BiLSTM**

Bidirectional LSTMs are an extension of LSTMs that can improve model performance on sequence classification problems. In problems where all time steps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. This can provide additional input context to the network and result in faster and even fuller learning on the problem. It involves duplicating the first periodic layer in the network so that there is now two layers' side-by-side, then providing the input sequence as-is as input to the first layer and providing a reversed copy of the input sequence to the second layer. The use of sequence bi-directionally was initially justified in the domain of speech recognition because there is evidence that the input context of the whole utterance is used to interpret what is being said rather than a simple interpretation. The use of bidirectional LSTMs may not make sense for all prediction problems but can offer benefits in terms of better results to those domains where it is appropriate.

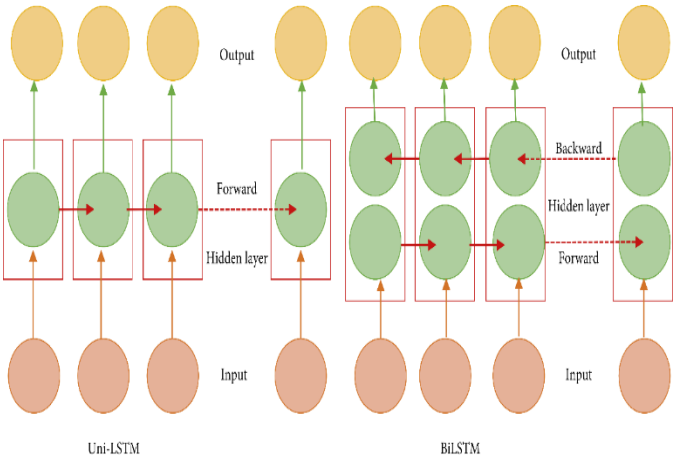
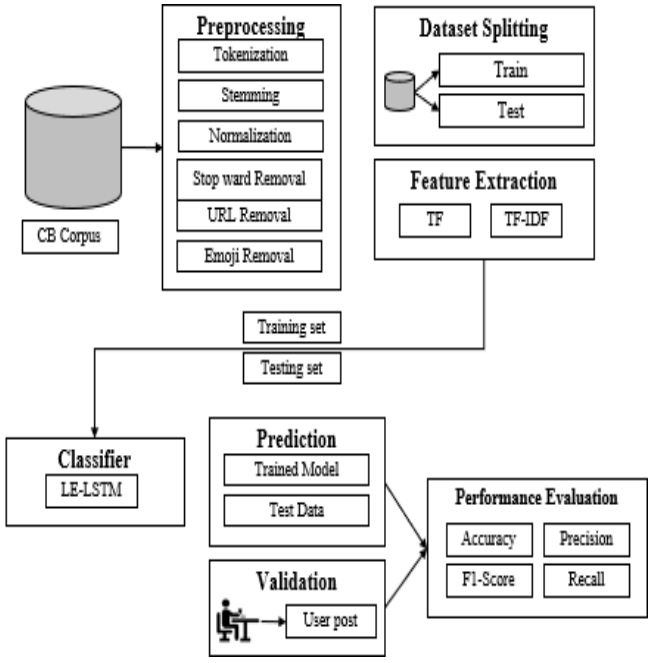


Diagram - BiLSTM

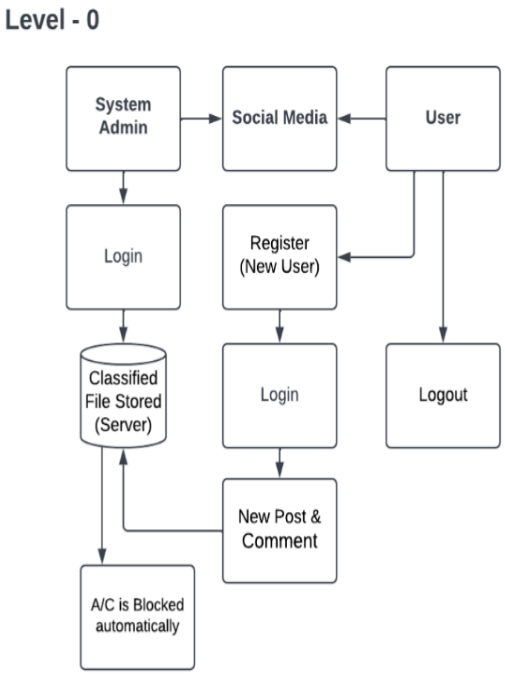
Bidirectional LSTM (BiLSTM) is a recurrent neural network used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. It's also a powerful tool for modelling the sequential dependencies between words and phrases in both directions of the sequence. In summary, BiLSTM adds one more LSTM layer, which reverses the direction of information flow. Briefly, it means that the input sequence flows backward in the additional LSTM layer. Then we combine the outputs from both LSTM layers in several ways, such as average, sum, multiplication, or concatenation.

4. SYSTEM ARCHITECTURE :

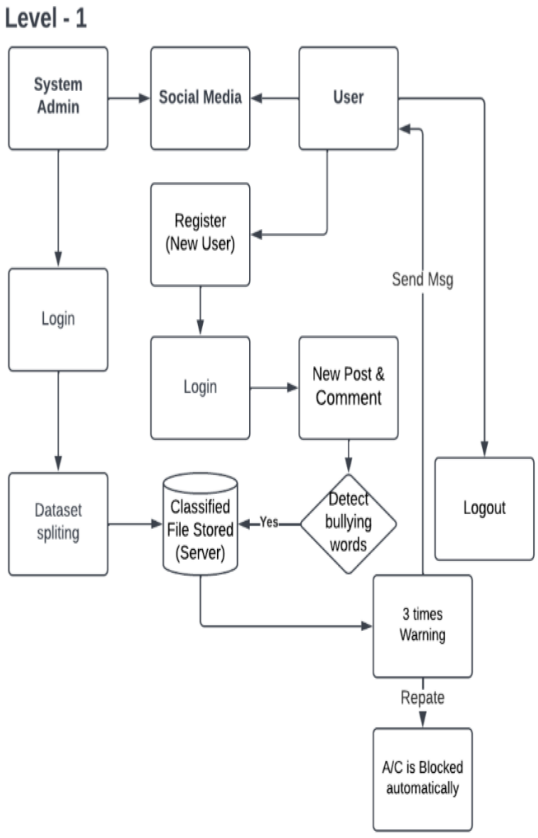


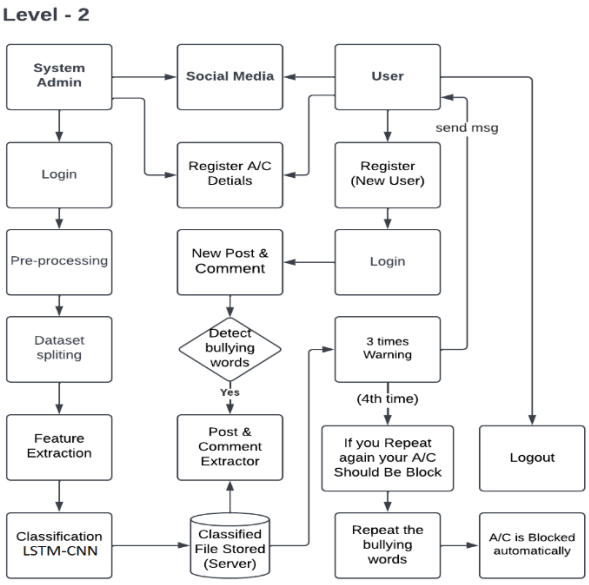
4.1. DATA FLOW DIAGAM:

4.1.1. Data Level 1:



4.1.2. Data Level 1:





5.ALGORITHMS

1. ALGORITHM-Text Preprocessing

Input: uncleaned tweets

Output: processed/cleaned tweet

Procedure:

1) For each tweet 2 uncleaned tweet
2) Delete all the special characters
3) Delete all single characters
4) Delete single characters from the begin
5) Substituting multiple spaces with a single space
6) Delete prexed `b'
7) transforming to Lowercase
8) Lemmatization

2.ALGORITHM - TF-IDF

TF-IDF algorithm split into two terms TF, which means how many words are in the current news.

TF (word)=number of repeated words appear in the document/total number of words in the document

where IDF refers to how necessary any terms are in all news. IDF gave a score to words.

IDF (word) = log (total amount of documents)/number of document where the word appears

Input:

1) D: Tweet
2) T: the unique term in all tweets

Output: weight matrix

Procedure:

1) For each ti €T do
2) For each tweets € do
3) Wij = number of appearances of term ti in tweets di
4) End for of tweets
5) End for of term

3.ALGORITHM - IDF

Input:

1) T: the unique term in all tweets
2) D: tweets
3) weight matrix from TF step

Output:

TF-IDF weight for each term

Procedure:

1) for each term ti € T do
2) for each tweet di € E
3) If TFij not equal zero, then EFi ++
4) End for of tweet
5) Idf - log(E/EFi)
6) End for of term
7) For each term ti do
8) For each tweetdj € E do
9) Tf- IDF = TFij *IEFi
10) End for of Tweet
11) End for of term

4.ALGORITHM-BiLSTM

Input: Training Twitter Tweets Dataset 'tnRSX', Training bullying/Nonbullying Values 'tnINY', Testing

Output: Accuracy ('Acc_total')

1. procedure BiLSTM MODEL (tnRSX , tnINY)
Hyperparameters with Arguments
2. batchsize = 32; epochs = 1; filters = 2; pool_size = 2; verbose = 2; N_EPOCH=7
3. max_features = 2000; embed_dim = 128; classes=2; input_length=3382; units=100
Build a Deep Learning Model
4. model = Sequential ()
Embedding Layer
5. model.add (Embedding (max_features, embed_dim, input_length))
Convolutional Layer
6. model.add (Conv1D (filters, kernel_size=3, padding='same', activation='relu')
Maxpooling Layer
7. model.add(MaxPooling1D(pool size))
BiLSTM Layer
8. model.add(LSTM(units))
Softmax Layer
9. model.add(Dense (classes, activation='softmax'))
Compile Function
10. model.compile (loss = 'binary_crossentropy', optimizer = 'adamax', metrics= [accuracy])
Model Summary
11. print (model.summary( ))
Fit a Model
12. for all epochs in (1: N_EPOCH) do
13. model.fit(tnRSX , tnINY , epochs, validation_data (ttRSX , ttINY, batch_size=batchsize))
#Model Evaluation
14. Acc = model.evaluate (ttRSX , ttINY, verbose, batch_size = batchsize)
15. Acc_total.append(Acc)
16. End for
17. return Acc_total
18. End Procedure

6.TESTING

In this phase of methodology, testing was carried out on the several application modules. Different kind of testing was done on the modules which are described in the following sections. Generally, tests were done against functional and non-functional requirements of the application following the test cases. Testing the application again and again helped it to become a reliable and stable system.

1.Usability Testing

This was done to determine the usability of the application that was developed. This helped to check whether the application would be easy to use or what pitfalls would the users come through. This was used to determine whether the application is user friendly. It was used to ascertain whether a new user can easily understand the application even before interacting with it so much. The major things checked were: the system flow from one page to another, whether the entry

points, icons and words used were functional, visible and easily understood by user.

2.Functional Testing

Functional Testing is defined as a type of testing which verifies that each function of the software application operates in conformance with the requirement specification. This testing mainly involves black box testing and it is not concerned about the source code of the application. Functional tests were done based on different kind of features and modules of the application and observed that whether the features are met actual project objectives and the modules are hundred percent functional. Functional tests, as shown in the following Table-1 to Table-5, were done based on use cases to determine success or failure of the system implementation and design. For each use case, testing measures were set with results being considered successful or unsuccessful. Below are the tables which are showing some of the major test cases along with their respective test results.

TABLE 1: Signup/Registration Test Case

Identifier	Test Case-1
Test Case	Signup
Description	To register new account in the application.
Pre-requisite	1) Username and email must not exist previously.
Test procedure	1) Select Sign Up from the menu. 2) Fill in username, email, and password and retype password accordingly. 3) Click on Sign Upbutton
Expected Result	1) User can register to the application successfully. 2) Username, email and password stored in the user table in the database.
Pass/Fail	Pass

Table 2: Login Test Case

Identifier	Test Case-2
Test Case	Login
Description	To login new account in the application
Pre-requisite	1) Registration must be done previously.
Test procedure	1) Select Log In from the menu. 2) Fill in username and password accordingly. 3) Click on Log In button.
Expected Result	1) User can login to the application successfully. 2) User should access the application features which are allowed
Pass/Fail	Pass

7. RESULTS AND DISCUSSION

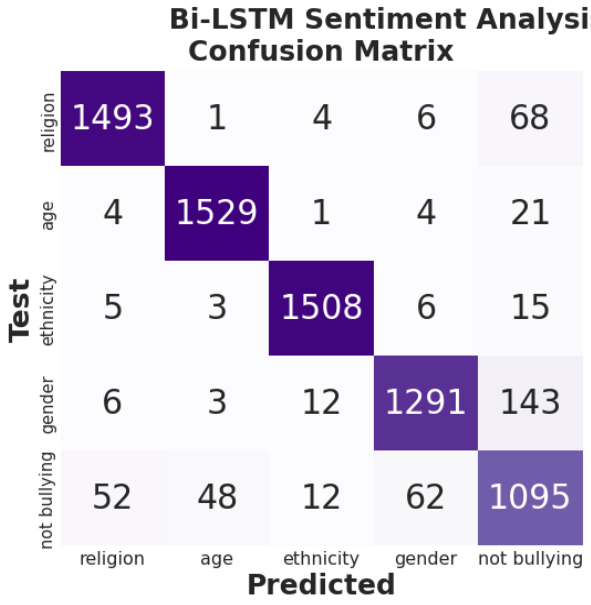
7.1. Evaluation Metrics

7.1.1. Confusion Matrix

The detection of spam emails can be evaluated by different performance measures. Confusion Matrix is being used to visualize the detection of the emails for models Several measurements are used for performance evaluation of classifiers like accuracy, precision, recall, and f-score. These measurements are computed by a confusion matrix, which is composed of four terms. Confusion matrix can be defined as below:

- True positive (TP): are the positive values correctly classified as positive.
- True Negative (TN): are the negative values correctly classified as negative.
- False Positive (FP): are the negative values incorrectly classified as positive.
- False Negative (FN): are the positive values incorrectly classified as negative.

For the performance evaluation of our proposed model, we use the following metrics.



7.1.2. Accuracy

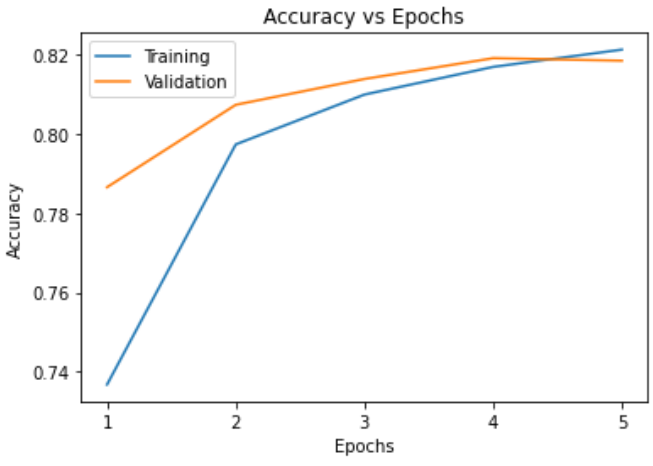
The accuracy measure is the ratio of the number of bully users detected to the total number

of bullies. It does not perform well with imbalanced data sets

$$\text{AccuracyCM} = \frac{\text{\# of detected bullies}}{\text{total number of bullies}}$$

training score :0.991249719542293

testing score :0.979372197309417



7.1.3. Precision

Precision is evaluation metrics used in binary classification tasks.Precision is the measure of exactness.

$$\text{Precision} = \frac{\text{\# of true bullies detected}}{\text{total number of detected users}}$$

In simple terms, high precision means that an algorithm returned substantially more bully users

7.1.4. Recall

The recall is a fraction of the predicted correctly classified applications to the total number of applications classified correctly or incorrectly. Recall is the measure of completeness.

$$\text{Recall} = \frac{\text{\# of true bullies detected}}{\text{total number of true bullies}}$$

whereas high recall means that an algorithm returned most of the bullies.

7.1.5.F-Score

F-score is the harmonic mean of precision and recall. It symbolizes the capability of the model for making fine distinctions. F1 Measure is the harmonic mean between precision and recall. The range for F1 is [0, 1]. It measures how many bullies are identified correctly and how robust it is. Mathematically, it can be expressed as

F1-Score=2*(Precision*Recall)/ (Precision + Recall)

F1 Measure attempts to find a balance between precision and recall. The greater the F1 Measure, the better is the performance of our approach.

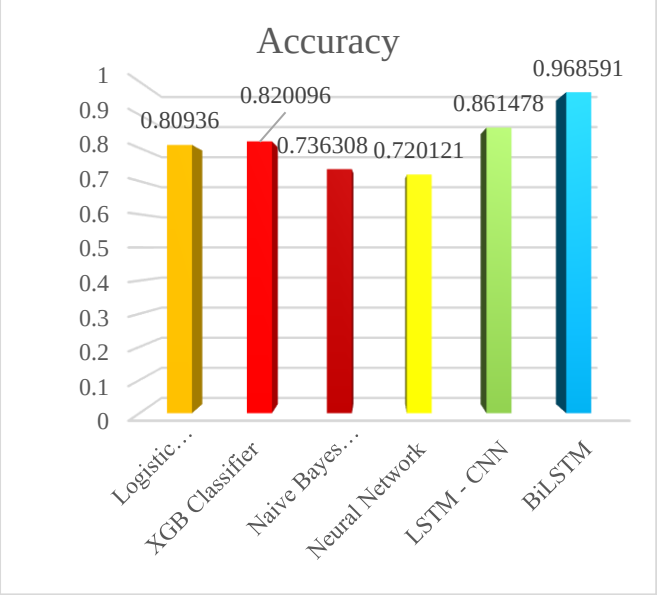
7.1.6. Results and Analysis

The results and comparisons of different classifiers after data training and testing are presented in this section. We gathered 49799tweets from the online resource ‘kaggle’ and translated them into English using the python library Google trans, which uses the Google Translate Ajax API. 42797tweets were used to train various ML and DL models. One seven thousand tweets were used for testing in order to quantify accuracy and assessment metrics. As explained about evaluation measures in chapter 9, we have evaluated accuracy, precision, recall, and f-measures that are evaluation measures measured using LR, XGBM and Naive Bayes, LSTM-CNN and BiLSTM. Finally, using various graphs, a comparison of models is presented below. The findings in Table 4 show that the deep learning algorithm (BiLSTM) is a stronger method for detecting cyberbullying tweet classification, with high accuracy of 98.4%.

Table 4: Accuracy of different models.

No	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.809360	0.816488	0.809360	0.812462
1	XGB Classifier	0.820096	0.829113	0.820096	0.823191
2	Naive Bayes Classifier	0.736308	0.723307	0.736308	0.708342
3	Neural Network	0.720121	0.808795	0.752411	0.778392
4.	LSTM - CNN	0.861478	0.895267	0.827423	0.852364
5.	BiLSTM	0.968591	0.986842	0.958647	0.984567

In the mentioned Table 4, we have compared the accuracy of four different ML and DL models. We can see that the DL model (BiLSTM) is the most accurate among all the models, but it takes a long time to train. ML models like LR, XGB and Naive Bayes are around the same accuracy percentage lower than LSTM/CNN and BiLSTM, which is also a DL model and has the lowest accuracy percentage. Figure 12 shows accuracy comparison of ML and DL models.



8.CONCLUSION

Cyberbullying is the harassment that takes place in digital devices such as mobile phones, computers and tablets. The means used to harass victims are very diverse: text messages, applications, social media, forums or interactive games.One of the things that complicates these types of situations that occur through the Internet, is the anonymity this environment allows. Since this facilitates cyberbullying can cover almost all areas of the victim's life, that is: educational environment, work, social or loving life. When the identity of the harasser is not known, even if the facts are reported, in many cases it is not enough to open an investigation, identify it and pay for the crime committed. This project proposed a deep learning model Bidirectional Long Short Term Memory (BiLSTM. Thus, this project has designed a method of automatically detecting the Cyberbullying attack cases. Identifies the messages or comments or posts which the BiLSTM model predicts as offensive or negative then it blocks that person id,then the admin can create automated reports and send to the concern department.Experiments are conducted to test three machine learning and 2 deep learning models that are; (1) GBM, (2) LR, (3) NB, (4) LSTM-CNN and (5) BiLSTM. This project also employed two feature representation techniques Tf and TF-IDF. The results showed that all models performed well on tweet dataset but our proposed BiLSTM classifier outperforms by using both TF and TF-IDF among all. Proposed model achieves the highest results using TF-IDF with 96% Accuracy, 92% Recall and 95% F1-score.

8.1. Future Scope

For the present, the bot works for Twitter, so it can be extended to various other social media platforms like Instagram, Reedit, etc. Currently, only images are classified for NSFW content, classifying text, videos could be an addition. A report tracking feature could be added along with a cross-platform Mobile / Desktop application (Progressive Web App) for the Admin. This model could be implemented for many languages like French, Spanish, Russian, etc. along with India languages like Hindi, Gujarati, etc.

9.SYSTEM SPECIFICATION

9.1. Hardware specification :

- Processors: Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores, 2 threads per core), 8 GB of DRAM
- Disk space: 320 GB
- Operating systems: Windows® 10, macOS*, and Linux*

9.2 Software specification

- Server Side : Python 3.7.4(64-bit) or (32-bit)
- Client Side : HTML, CSS, Bootstrap
- IDE : Flask 1.1.1
- Back end : MySQL 5.
- Server : WampServer 2i
- BC DLL : PyChain, Node Package Manager, Virtualenv, Block chainhash

REFERENCES:

1. A. S. Srinath, H. Johnson, G. G. Dagher and M. Long, "BullyNet: Unmasking cyberbullies on social networks", IEEE Trans. Computat. Social Syst., vol. 8, no. 2, pp. 332-344, Apr. 2021.
2. Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa and M. Wroczynski, "Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection", Inf. Process. Manage., vol. 58, no. 4, Jul. 2021.
3. N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, et al., "Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking", Math. Problems Eng., vol. 2021, pp. 1-12, Feb. 2021.
4. R. R. Dalvi, S. B. Chavan and A. Halbe, "Detecting a Twitter cyberbullying using machine learning", Ann. Romanian Soc. Cell Biol., vol. 25, no. 4, pp. 16307-16315, 2021.
5. N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, et al., "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification", Comput. Electr. Eng., vol. 92, Jun. 2021.
6. A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in Arabic tweets using deep learning", Multimedia Syst., Jan. 2021.
7. Y. Fang, S. Yang, B. Zhao and C. Huang, "Cyberbullying detection in social networks using bi-GRU with self-attention mechanism", Information, vol. 12, no. 4, pp. 171, Apr. 2021.
8. B. A. Talpur and D. O'Sullivan, "Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter", Informatics, vol. 7, no. 4, pp. 52, Nov. 2020.
9. A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan and M. Prasad, "Identification and classification of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting" in Neural Information Processing, Cham, Switzerland:Springer, vol. 1333, pp. 113-120, 2020.
10. C. Iwendi, G. Srivastava, S. Khan and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures", Multimedia Syst., 2020.
11. L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "XBully: Cyberbullying detection within a multi-modal context," in Proc. 12th ACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 339–347.
12. C. Van Hee et al., "Automatic Detection of Cyberbullying in Social Media Text." 2018.
13. M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. Shalin, and A. Sheth, "A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research," pp. 33–36, 2018.
14. A. H. Alduailej and M. B. Khan, "The challenge of cyberbullying and its automatic detection in Arabic

text," 2017 Int. Conf. Comput. Appl. ICCA 2017, pp. 389–394, 2017.

15. A. Power, A. Keane, B. Nolan, and B. O. Neill, "A lexical database for public textual cyberbullying detection," Rev. Lenguas Para Fines Específicos, vol. 2, pp. 157–186, 2017.
16. M. Drahošová and P. Balco, "ScienceDirect The analysis of advantages and disadvantages of use of social media the analysis of advantages and disadvantages of use of social media in European Union in European Union," Procedia Comput. Sci., vol. 109, pp. 1005–1009, 2017.
17. R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in Proc. 17th Int. Conf. Distrib. Comput. Netw., Jan. 2016, pp. 1–6.
18. V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2016, pp. 884–887.
19. A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, Aug. 2015, pp. 280–285.
20. P. Galán-García, J. G. De La Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," Logic J. IGPL, vol. 24, no. 1, pp. 42–53, 2015.