IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Detecting Twitter Cyberbullying Using Deep learning

VISHRUTHA S N
Department of Computer Science Science Dayananda Sagar College of Engineering Engineering
Bengaluru, India

J VIMALA DEVI
Department of Computer Dayananda Sagar College of
Bengaluru, India

***Abstract*— In today's online world, cyberbullying is an escalating problem that can trigger deep emotional pain and isolation from others. There is an increasing necessity to regulate content shared on online platforms. Cyberbullying material can spread quickly across online platforms and may remain visible for a long time, sometimes without ever being removed. Such persistent exposure to harmful content can take a serious toll on the mental, emotional, and psychological health of young people. In extreme situations, ongoing harassment in digital spaces can contribute to thoughts of self-harm or suicide. The proposed system utilizes Long Short-Term Memory (LSTM) networks to assess user-generated text and compute a bullying likelihood for every sentence. This system monitors user activity in real time and decreases a reputation score whenever harmful or bullying content is identified. Once the score falls below a set limit, the user is automatically prevented from continuing interactions on the platform. By merging deep learning with a reputation- based penalty method, the design works to curb cyberbullying while ensuring moderation remains both fair and proactive. The approach is adaptable, efficient, and capable of supporting healthier online spaces.**

Keywords*— *cyberbullying;* *Deep learning; LSTM;

I. INTRODUCTION

Cyberbullying is the use of digital communication systems, such as social media, messaging apps, and other online channels, to threaten, mistreat, or deliberately cause harm to someone. It often involves insults, humiliation, or other forms of targeted abuse that can cause both emotional and physical distress, and in severe cases, may lead to self-harm or suicidal behaviour. One of the biggest challenges in addressing cyberbullying is identifying it at an early stage so that timely action can be taken. People may engage in cyberbullying for various reasons, such as seeking attention, trying to gain social status, or attempting to show control over others. In some cases, individuals may not fully understand the harm their actions can cause. Guidance and oversight from parents or guardians can play a significant role in reducing such behaviour, particularly by monitoring children's online activity, smartphone usage, and the kind of material they engage with on social media.

In the modern digital environment, no age group or community is entirely safe from becoming a target — even older adults can be affected. Disagreements over opinions can sometimes escalate into personal attacks, where one person intentionally tries to harm another. Such behaviour, whether delivered through spoken words, written messages, physical acts, or public humiliation, falls under the umbrella of bullying. The following figure shows the Insights into global patterns of social media usage in 2025.

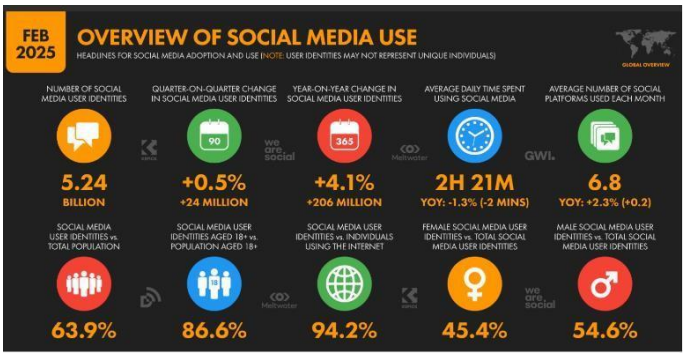


Figure 1: Insights into global patterns of social media usage in 2025

Many countries have introduced rules and policies aimed at reducing and controlling cyberbullying. However, because this behaviour can take place across numerous social media platforms, detecting it in practice can be challenging. Cyberbullying is an intentional behaviour intended to insult, threaten, or humiliate a person. Such practices may include dissemination of fake news, sharing of personal photos or videos without consent, obscene content, or any other harassment, online. The most popular of the manifestations is the use of abusive or derogatory textual material. There are a number of methodological and technological interventions that have been investigated over the years in order to combat this phenomenon. The Web 2.0 has brought a radical shift in digital communication where the way of communicating and how relationships are made are changed. Not only have these technological inventions made people to be connected, but they have also offered channels through which evil acts can be reached, particularly among the teens who are highly active in social sites. The cyberbullying problem must be addressed

holistically and, in this case, the most significant variables are automated detection and prevention. The above methods have been based predominantly on the traditional machine-learning algorithms but the new advancements in the deep-learning algorithms have significantly improved the detection accuracy. In addition, some of these models have been designed as cross- platform, thus proving to be flexible in other social-media environments and data sources. In this case, we apply these models to a novel dataset gathered on Twitter, thereby questioning their generalizability and functionality in cross- platform environments. This comparative analysis can be used to assess the effectiveness of deep-learning models in parallel to the traditional ML methods and explain the extent to which models trained on one dataset can generalize to another dataset.

II. LITERATURE SURVEY

Our work also reviewed earlier studies that explored different processing methods. The aim was to understand how these approaches can help detect negative behaviour and encourage more positive interactions in online comments.

Darko Tosev, Sonja Gievska et al. [1] participated in research on social media hate speech detection. SVM, RF, and Logistic Regression have been used in a multi-level layered collaborative learning technique, together with a variety of sparse and dense feature representations. Future usage of pre- trained vector embeddings and other machine learning techniques may be necessary due to the ensemble model's low accuracy results.

Mona Khalifa et al. [2] suggested a brand-new hybrid mutation method based on Darwinian principles termed the Genetic Programming (GP) model to address the binary text classification problem of social media cyberbullying detection. The suggested methodology has not been applied to multi- classification issues; instead, it handles text classification as a categorization of binary issues.

Shreelakshmi K, Premjith B et al. [3] offered a ML system for detecting hate speech in Hindi and English. The algorithm combines text from social media with faster, better feature representation using SVM. Text characteristics are used to produce the RBF classifiers. a little, binary data collection that is solely separated into two categories: hate and non-hatred.

Zafer Al-Makhadmeh et al. [4] offered a hybrid approach to social media hate talk classification that combines natural language processing (NLP) and machine learning (ML). The study offers a way for categorizing hate talk from social media sites by combining ML and NLP approach. Before starting the hate talk recognition procedure, steaming, token splitting, character and infection eradication, and crowd flowering are done after the hate speech has been collected. Semantic, sentiment, unigram, and pattern characteristics are NLP features that are extracted after the system analyses Tweets for sentences and words. Then the data gathered is investigated using a killer natural language processing optimization ensemble deep neural network model (KNLPEDNN). Classification of Cyberbullying detection is performed over an English dataset and not used over a multiple language dataset.

Muhammad Okky Ibrahim et al. [5] presented research that employs ML with Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifiers as the data transformation approach to address multi- label text categorization for abusive language and cyberbullying detection. According to error analysis, the unbalanced dataset is most likely the source of many false- negative mistakes.

Aditya Gaydhani et al. [6] suggested a method for using a ML to detect toxic language on publicly accessible Twitter datasets. Using SVM, NB, and Logistic Regression as classifier models, a comparison of the outcomes was achieved. After adjusting the hyperparameters, Logistic Regression outperforms the other three models. Due to the lack of other samples of harmful language that would not contain hateful phrases, it was observed that 4.8% of the inflammatory tweets were incorrectly labeled as hateful.

M. Ali Fauzi et al. [7], suggested the most effective technique for detecting cyberbullying in Indonesian. On the Twitter hate speech dataset, they used two techniques, hard voting and soft voting, along with five standalone classification algorithms, including Naïve Bayes, (K-NN) K-Nearest Neighbors, Maximum Entropy, Random Forest, and SVM. Using an ensemble of the three best yields the greatest results; these methods have outperformed nearly all of the stand-alone classifiers. demonstrating how the system's performance risk of using a subpar classifier can be reduced by employing the optimal approach. Despite using ensembles of feature sets and BOW, the results did not much improve. Nabiila.

Adani Setyadi et al. [8] employed a technique to categorize hate talk components within a text so that hate speech may be identified later. utilizing a backpropagation algorithm-optimized Artificial Neural Network technique. A supervised learning training technique using a target input pair intended for multi-layer perceptron operation is part of the backpropagation algorithm. This algorithm, as its name suggests, carries out two steps of computation: an advanced calculation to determine the error between the target and actual output. To adjust the weights on all of the current neurons, the error is propagated using a backward calculation. The procedures involved in training a multi-layer perceptron with two hidden layers using the backpropagation technique. The only restriction was the Boolean classification of tweets as either hateful or non-hateful.

Kelvin Kiema Kiilu et al. [9] introduced a system formulates a method for identifying and classifying hate speech using content generated by hateful communities that identify themselves on Twitter. This article contrasted different

supervised machine learning algorithms of Naïve Bayes' to conduct sentiment analysis and hate tweet detection on Twitter. Aside from the predictive capability of the system for a specific tweet on being hateful or not, the system also produces a list of users who tend to post these frequently. This gives us an interesting overview on the pattern of use of hate- mongers concerning how they articulate bigotry, racism and propaganda. The problem solved is the restriction of Twitter API to commercial research wherein authorization is restricted.

Hajime Watanabe et al. [10] suggested approach automatically identifies the most prevalent unigrams and hate speech patterns, then classifies tweets into three categories: clean, offensive, and hateful. It also incorporates emotive and semantic aspects. Words that are typically used to offend, denigrate, or insult somebody are also strongly associated with hate. As a result, when we separate the class "offensive" from the binary divided into two classes, "hateful" and "offensive," tweet features classed as "Unigram" offer poorer accuracy.

Sanjana Sharma et al. [11] described how to use Naïve Bayes, SVM, and random forest classification to annotate Twitter data in accordance with an ontological classification of damaging speech based on the level of malicious intent. In order to provide a workable classification model for identifying harmful speech by presenting an annotated corpus of tweets that were categorized according to different levels of hate in this paper. We used our linear classification skeleton of harmful speech to identify three classes (Class I, Class II, and Class III) in the Twitter dataset. When employing Random Forest as the classification algorithm and the bag of words technique in the feature vector produced the highest accuracy of 76.42%.

Zhang, X., Luo, L., Fung, P., & Liu [12] using a deep learning approach that uses CNN and LSTM to detect cyberbullying on social media sites. The significance of feature extraction and the effects of unbalanced datasets are covered.

Gaur, M., & Ahuja, A [13] provides an overview of cyberbullying detection methods that includes both machine learning with social pedagogical measures. It discusses the importance of context and social factors in identifying cyberbullying incidents.

De Silva, D., Manogaran, G., & Lopez, D [14] concentrates on employing NLP methods to detect cyberbullying on Twitter. For better identification, it investigates feature engineering, topic modeling, and sentiment analysis.

Mishra, S., & Shen, B. [15] examines how deep learning techniques can be used to detect cyberbullying. The performance of several deep learning architectures, including CNN, LSTM, and Transformer models, is covered.

Sarker, A. H., & Yang, Y. [16] suggests using deep learning to identify instances of cyberbullying on Instagram. It emphasizes the examination of textual and visual content, highlighting the significance of multimodal analysis.

Hussain, M., & Al-Sarem, M.[17] examines how methods for detecting cyberbullying have changed over time and emphasizes how machine learning might help with this problem. It also describes the field's difficulties and potential paths forward.

III. METHODOLOGY

1. Proposed Solution:

Online bullying has remained a significant issue in contemporary society, and it tends to damage the psychological well-being of individuals, particularly the young generation that uses social media platforms. This project aims at developing a smart system that will be able to automatically identify such harmful behaviour with Long Short-Term Memory (LSTM) networks. The system verifies every sentence, assigns it a score depending on the likelihood that it is bullying, and logs all users with a reputation index. In case the score of a user decreases below a required threshold, the user will be restricted in their activity to prevent further adverse interactions. The system offers an active means to minimise online harassment using natural language processing (NLP), machine learning, and real-time monitoring. The principal objective is to create a more secure virtual environment in which individuals can socialise in a positive and respectful way.

2. System Architecture:

The proposed project presents a real-time cyberbullying detection system that can be used to encourage positive and respectful Internet communication. The system starts with the user-generated text that is captured using a web-based interface and sent to a Flask-based backend. Input text is then pre- processed to filter out undesired items and is ready to be analysed.
LSTM model receives an input of processed text and outputs a bullying probability score. According to this score, a reputation management module updates the ex of each user, and this is kept in a central database.
It is also a real time framework and will be used to give immediate feedback to the user interface and the monitoring elements so that the harmful content can be identified and addressed real time. This kind of integrated approach can be used to make the online space secure, beneficial and fruitful.

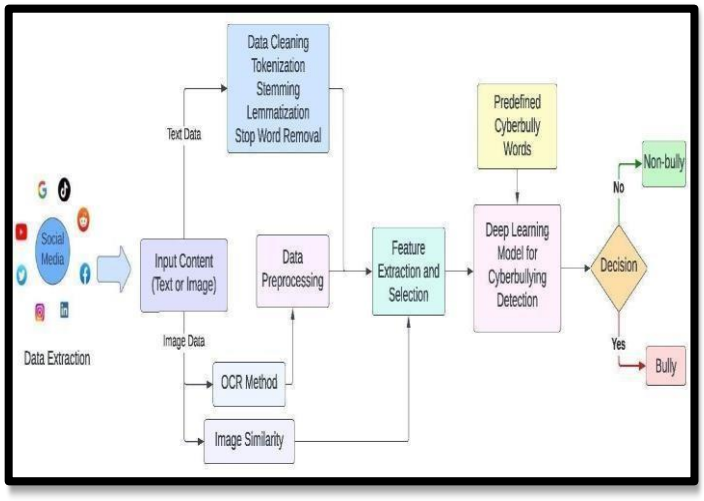


Figure 2: System Methodology

1. **Data Collection:**

Prepare a collection of text samples, e.g., chat messages, social media posts, or online conversations, each of them tagged to show the degree of cyberbullying found. It should be represented in the dataset by a combination of safe and harmful content to allow the LSTM model to learn how to identify patterns correctly.

2. **Preprocessing:**

Prepare the text by removing unwanted characters, unnecessary spaces, and commonly used stop words. Divide the sentences into small pieces (tokens) and convert them into a form that the LSTM model can understand. The likelihood of bullying is then given to each sentence, which assists the system in estimating the degree of offensiveness or safety of the content.

Example:

Input: "YOU ARE USELESS!!!" After cleaning: "you are useless" After removing stopword: "useless" After tokenization: ["useless"] After embedding: [345]

3. **Model Development:**

The system breaks down words in a message into sequence with a Long Short-Term Memory (LSTM) neural network. Since LSTMs are able to recall information on longer sentences, they may be utilized to recognize the offensive or abusive words better. The model is trained so that it provides bullying score on all sentences, which reflects the severity of the identified behaviour.

Example: "you are stupid" → 0.87 (high bullying likelihood)

If the user reputation score is less than the threshold reputation scores his/her interaction will not happen further

4. **Reputation Score System**:

It presents a reputation tracking system where all users start with a maximum score. Each time the LSTM model recognises a message as bullying, the score is reduced relative to the degree of the behaviour. Repeated violations have further effects of reducing the scores, and consistent positive activity has a gradual recovery of the score.

Formula to calculate Reputation score: user Info['score']/user Info['total'] *10

5. **Blocking Mechanism**:

A reputation score has been set in order to keep fairness within the platform. When a user scores below this limit as a result of recurrent bullying cases, the system will automatically limit his/her access and will not allow further interaction.

6. **Testing and Evaluation**:

New data sets are used to test the model. Various evaluation measures - such as accuracy, precision, and recall - are used to evaluate the performance of the system in identifying malicious content. In addition, the blocking mechanism is reviewed with respect to its ability to react effectively to repeated violations.

7. **Deployment**:

The presentation of a web application is presented, where Flask is used as the back-end, where the trained model is integrated. The real-time surveillance will be used to make sure that bad material will never be identified in time, and the automatic block feature will act immediately on the violators to keep the virtual world clean and safe.

Novel Contributions:

This study introduces numerous additional techniques compared to existing approaches:

1. **Integrating Deep Learning with Reputation Management :**
   Aside from various classification, the algorithm maintains dynamic reputation scores for each user.
2. **Real-Time Blocking Mechanism:**
   This feature ensures proactive protection when dangerous behaviour is verified and are blocked by this mechanism.

3. **Cross-Platform Adaptability:**
   Although proven on Twitter, the architecture is versatile and could potentially be applied across several social media or chat systems.
4. **Empirical Validation:**
   A comparison with regard to baseline machine learning models suggests that deep learning excels them for sequential text classification.

IV. RESULT:

Accuracy:

The detection model was based on LSTM, and the mode attained an accuracy of 98.25 percent. This percentage is an indicator of the percentage of sentences that are identified as either bullying or non-bullying. By contrast, the other models, including XG Boost (**71.35%),** Random Forest (**66.82%**), and SVC (**71.50%**), provided worse accuracy. The high accuracy of the LSTM model underscores its prowess in not only depicting sequential but also contextual information in text, thus it more dependable in identifying cyberbullying.

Model	Accuracy
LSTM	98.25%
SVC	71.50%
XG Boost	71.35%
Random Forest	66.82%

Table 1: Performance Evaluation

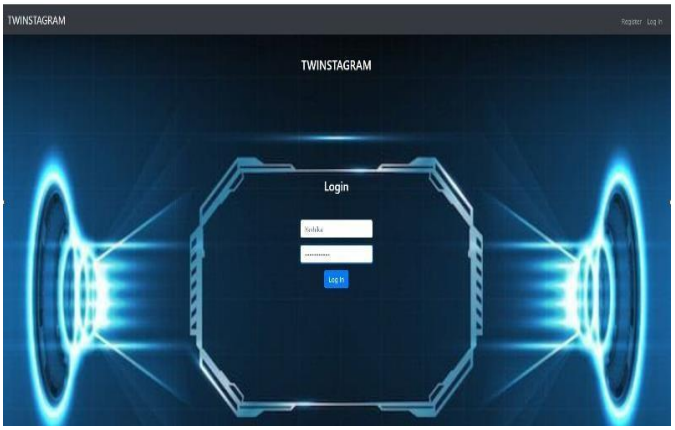


Figure 3: Login Page



Figure 4: Cyber Bullying Detection Using LSTM

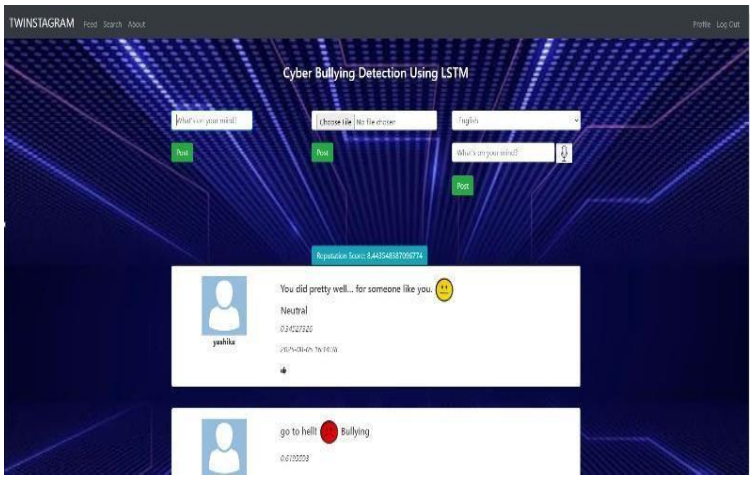


Figure 5: Detecting Cyber Bullying

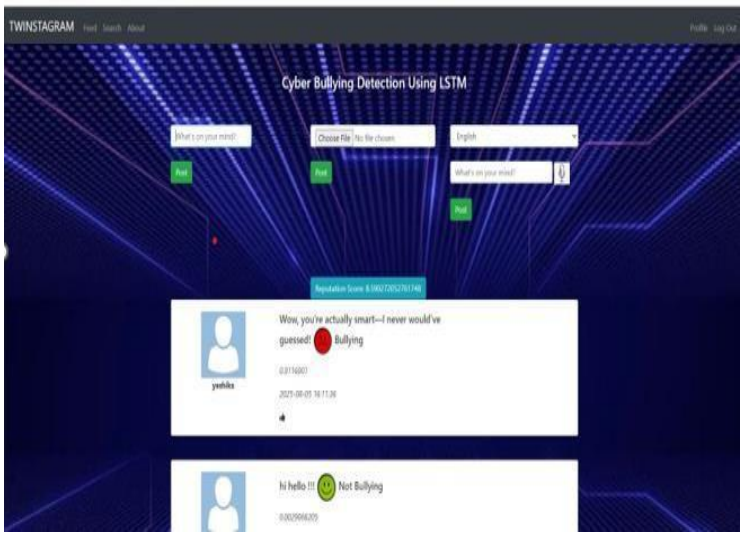


Figure 6: Detecting non-Cyber Bullying

Models Evaluated:

To ensure robustness, several models were tested:

1. **Support Vector Classifier (SVC):**
   Good for text but less effective for collecting sequential context. Accuracy is 71.50%.
2. **Random Forest:**
   Performs well with organized features but poorly with sequential data. Accuracy is 66.82%.
3. **XGBoost:**
   Provides a strong gradient-boosting baseline but lacks contextual awareness. Accuracy is 71.35%.
4. **LSTM (Proposed):**
   Designed for sequential language modelling. Accuracy: **98.25%.**

V. CONCLUSION:

The paper has proposed an automated method of identifying cyberbullying through training a Long Short-Term Memory (LSTM) model with a natural language processing view to understand the text of the users. The system also verifies all the sentences as possible bullying acts, and also logs the user activity by ranking the user on a reputation system. In such a manner, it could identify and restrict malicious interactions. This model achieved an accuracy of **98.25 per cent**, which is higher than older machine learning methods and demonstrates the importance of the context of the sentence. In general, the system is a rapid scaling method of supporting a more respectful and safe online discussion.

VI. FUTURE WORK:

The system can be improved in the future by training on larger and more diversified datasets, which include other languages, words of slang words, and diverse expressions of culture. Besides, other types of data, including images, video, or voice clips, might be included to identify more complicated instances of bullying. A second helpful measure would be to introduce explainable AI functionality; thus, users and moderators can clearly understand the reason behind the marking of specific messages. Lastly, the model will be directly tested on active social media platforms to be more practical and feasible and its long-term result will be observed.

REFERENCES

[1] Das, Kumar & Garai, Buddhadeb & Das, Srijan & Patra, Braja. (2021). Profiling Hate Speech Spreaders on Twitter-Notebook for PAN at CLEF 2021. Media,"

[2] M. K. A. Aljero and N. Dimililer, "Genetic Programming Approach to Detect Hate Speech in Social in IEEE Access, vol.9,pp.115115- 115125,2021,doi: 10.1109/ACCESS.2021.3104535.

[3]] K. Sreelakshmi, B. Premjith, K.P. Soman, "Detection of Hate Speech Text in Hindi-English code-mixed Data", Procedia Computer Science, Vol. No. 171, Page No. 737-744, 2020

[4] Al-Makhadmeh, Zafer, and Amr Tolba. "Automatic Cyberbullying detection using killernatural language processing optimizing ensemble deep learning approach." Computing 102, no. 2 (2020): 501-522.

[5] Ibrohim, Muhammad Okky, and Indra Budi. "Multi-label hate speech and abusive language detection in Indonesian twitter." In Proceedings of the Third Workshop on Abusive Language Online, pp. 46-57. 2019.

[6] Aditya Gayadhani, Vikrant Doma, Shrikant Kndre, Laxmi Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach", IEEE International Advance Computing Conference (2018), 2018.

[7] Fauzi, M. Ali, and Anny Yuniarti. "Ensemble method for Indonesian twitter Cyberbullying detection." Indonesian Journal of Electrical Engineering and Computer Science 11.1 (2018): 294-299.

[8] N. A. Setyadi, M. Nasrun and C. Setianingsih, "Text Analysis for Cyberbullying detection Using Backpropagation Neural Network," 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2018, pp. 159-165, doi: 10.1109/ICCEREC.2018.8712109.

[9] Kiilu, Kelvin & Okeyo, George & Rimiru, Richard & Ogada, Kennedy. (2018). "Using Naïve Bayes Algorithm in detection of Hate Tweets. International Journal of Scientific and Research Publications" (IJSRP). 8. 10.29322/IJSRP.8.3. 2018.p7517.

[10] Rui Zhao, Kezhi Mao "Cyber Bullying Detection based on Semantic - Enhanced Marginalized Denoising Auto-encoders". IEEE Transaction on Affective Computing, 2015.

[11] Elaheh Raisi,Bert Huang "Weakly Supervised Cyberbullying Detection with Participant Vocabulary Consistency" Social Network Analysis and Mining, May 24,2018.

[12] Peng Zhou,Wei Shi,Jun Tian,Zhenyu Qi,Bingchen Li,Houng Wei,Hao,Bo Xu "Attention- based Bi-directional Long ShortTerm Memory Network for Relation Classification" proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ,pages 207 212,August 12,2016.

[13] Nitish Srivastava, Geoffrey Honton,Alex Krizhevsky,Iiya Sutskever,Ruslan Salakhutdinov "Dropout: Simple way to Prevent Neural Networks from Overfitting" Journal of Machine Learning Research 1929- 1958,2015

[14] Alexis Conneau,Holger Schwenk,Yann Le cun "Very Deep CNN for Text Classification" Association for Computational Linguistics, Volume1, pages 1107-1116,7 April 2017.

[15] MS.Snehal Bhoir,Tushar Ghorpade,Vanita Mane "Comparative Analysis of Different Word Embedding Models" IEEE,2017.

[16] Elaheh Raisis, Bert Huang "Cyberbullying Detection with Weakly Supervised Machine Learning" International Conference on Advances in Social Networks Analysis and Mining IEEE/ACM,2017.

[17] Haipeng Zeng,Hammad Haleem,Xavier Plantaz,NanCao and Huamin Qu "CNN Comparator: Comparative Analytics of CNN" arXiv,15 Oct,2017