ENHANCING CYBERBULLYING DETECTION ON SOCIAL MEDIA USING NLP AND MACHINE LEARNING TECHNIQUES

G. Jayagopi
Department of Computer Science and Engineering
Mother Theresa Institute of Engineering and Technology, Palamanar, India
jayagopi0075@gmail.com

CB. Sumathi
PG and Research Department of Mathematics
Marudhar Kesari Jain College for Women, Tamil Nadu, India
c.bsumathi@yahoo.in
https://doi.org/10.34293/9789361639715.shanlax.ch.026

Abstract

The exponential growth of social media platforms, fuelled by the widespread accessibility of the Internet, has transformed the way individuals communicate and interact in the 21st century. While these platforms have enhanced social connectedness and fostered new opportunities for sharing information, they have also inadvertently facilitated an increase in harmful online behaviours, including cyberbullying, cyberstalking, and various forms of digital harassment. These negative interactions pose significant risks, particularly to vulnerable groups such as women and children, who often suffer severe psychological and physical consequences. In extreme cases, cyberbullying has been associated with suicidal ideation and mental health crises, underscoring its profound societal impact. The global prevalence of such incidents – ranging from the dissemination of private messages and spreading of malicious rumours to sexually explicit comments – has heightened public awareness and concern. This has consequently spurred a growing body of research aimed at developing effective strategies to detect and mitigate cyberbullying. In response to this urgent need, the present study explores the integration of natural language processing (NLP) techniques with machine learning algorithms to build a robust framework for identifying cyberbullying and abusive posts across social media platforms. We evaluate the performance of four distinct machine learning classifiers using two widely adopted feature extraction methods: Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). By systematically comparing these approaches, this research aims to advance the accuracy and reliability of automated cyberbullying detection systems, contributing valuable insights toward safer online environments.

Keywords: *Cyberbullying, Artificial Intelligence, Natural Language Processing, Machine Learning, Social Media, Online Abuse, Text Analysis*

Introduction

The term social media describes the wide spectrum of internet sites and communities which allow users to form, exchange and interact with any form of digital content such as text, pictures, video, and documents [1]. These platforms can be accessed mainly through laptops, smartphones, tablets, and other connected devices, and communication is real time and global. Popular social networks Facebook, Twitter, Instagram, Tik Tok, and others are now an inseparable part of everyday life that they affect the social, professional, and cultural relationships of people. In addition to informal communication, social media is being used more in the education sector [2], business world [3], as well as in the charitable arena [4]. This mass adoption has also helped the economy of

the world in many ways since it has generated a lot of employment opportunities in various fields which pertain to technology, marketing and content creation [5].

Regardless of all positive aspects that social media brings to the lives of users, like the development of social interactions, the ability to share information in real time, and creative outlets, certain risks and challenges are present. Among the most disturbing ones, the emergence of unethical and harmful behaviours, which may inflict emotional pain and ruin personal reputations, should be mentioned. Cyberbullying has proved to be a serious societal issue among these harmful behaviours. Cyberbullying can also be termed as cyber harassment; it means the use of electronic devices to harass, intimidate or humiliate other people. Such bullying goes beyond the physical location and may take place in different mediums like social media updates, personal messages, email, and internet forums.

With the development of digital communication technologies, the number of cyberbullying cases has become frightening, and this tendency is observed mostly in adolescents and teenagers because this population group is highly susceptible to the negative impact of cyberbullying. Studies show that almost 50 percent of American adolescents have undergone some form of cyberbullying [6]. The psychological, emotional and sometimes even physical burden imposed on the victims can be heavy and makes them anxious, depressed, socially withdrawn and in some extreme cases, the victims develop suicidal thoughts and attempts [7][8]. Cyberbullying is pervasive and therefore requires proactive action that will help detect and intervene on it to reduce the harmful effects of cyberbullying.

Responding to this threat, this paper offers a machine learning-based solution to the automated identification of cyberbullying in textual content. Using the methods of natural language processing (NLP), it is our goal to develop a useful classification mechanism that will be able to differentiate between cyberbullying-related and non-cyberbullying posts. We plan to use our methodology of assessing various machine learning algorithms trained and tested on actual data collected in the social media including Twitter and Facebook. In order to encode textual data as a numerical value, we select two widely used feature extraction methods, namely Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Experimental findings of our study have shown that TF-IDF is always superior to BoW in identifying relevant textual characteristics and out of the classifiers studied, the Support Vector Machines (SVM) have been found to be the best in terms of accuracy and strength.

Related Works

Cyberbullying detection on machine learning has developed and many studies are exploring different methodologies. Initial methods tended to use the conventional supervised learning models. Indicatively, a bag-of-words (BoW) approach was applied in identifying emotional and contextual features of text with an accuracy of 61.9 only [9]. Extending this, Ruminati et al. [10] at MIT took a Support Vector Machine (SVM) to identify cyberbullying in YouTube comments by adding social factors and commonsense reasoning to 66.7 percent accuracy. Reynolds et al. [11] proposed a language-based model based on the

decision trees and the instance-based learning with the accuracy of 78.5. A further work [12] revealed usefulness of richer linguistic characteristics, including personality traits, emotions and sentiment in enhancing performance in detection.

With the advancement of the complexity of online communication, it became more efficient to resort to deep learning techniques as the alternatives. Badjatiya et al. [14] used deep neural network (DNN) models in hate speech detection, which is a similar task, and others used convolutional neural networks (CNNs) with word representations to model semantic relationships more effectively [15]. These developments allowed models to see subtle patterns that were not available to the more traditional methods.

Understanding that the behaviour of cyberbullying is usually not limited to text only, researchers have shifted to multi-modes. The system, which was introduced by Cheng et al. [16], XBully, is used to convert multi-modal data (text, images, social interactions) into a heterogeneous network to study cross-modal relations. On a similar note, Wang et al. [17] created a multi-modal cyberbullying detection model that accommodates image, video, text, and temporal data, based on hierarchical attention networks to model the social session and multimedia content.

The models of hybrid neural networks with LSTM and CNN layers have also proven to be promising. Buan et al. [18] suggested a stacked-layer model with a mechanism of Support Vector Machine-like activation based on L2 regularization and hinge loss, which is better at enhancing generalization of the model. In overcoming the lack of labeled data, Raisi et al. [19] introduced a semi-supervised co-training system based on the keywords provided by experts and the social network connectivity, minimizing the co-training and the weak supervision losses.

There has also been attention of platform-specific approaches. The model by Al et al. [20] was Twitter-centric, with behavioural characteristics and tweet text, with a high F1-score of 0.936 and ROC-AUC of 0.943. Singh et al. [21] also paid special attention to the role of integrating visual data, demonstrating that the accuracy of cyberbullying detection can be greatly increased with the use of image features. Also, new methods such as Fuzzy Fingerprints are explored to analyse textual patterns in social media content [22].

Although these developments are made, there are still enormous gaps. Firstly, most systems are unable to conduct real time detection which is a vital requirement needed to intervene in time. Second, these models do not easily generate cross-platform transfers because data formats, language use, and user behaviour are different. Third, lack of extensive, label data with multi-modes restricts the training and benchmarking of models that would be used to understand cyberbullying as it is complex. In addition, although visual data integration is on the rise, existing systems continue to waste video and image context-although the usage of memes, screenshots and videos in the context of bullying behaviour is increasing. Ethical issues like privacy, equity and consequences of false positives are also not addressed. These issues allow identifying the necessity of additional research aimed at creating scalable, explainable, and ethically responsible cyberbullying detection algorithms that would be able to adapt to changing online behaviours.

Bullying Detection Model

The suggested framework of cyberbullying detection will include two components that will be inseparable: the use of Natural Language Processing (NLP) to preprocess and extract features out of the text, and then the use of Machine Learning (ML) algorithms to classify the features. First, the data sets with examples of bullying and non-bullying messages on social media, including Facebook and Twitter, are gathered. NLP tricks are then used to get the raw text data into shape so that it can be classified by machine learning with ease.

Data Preprocessing and Feature Engineering

There are many unnecessary symbols, punctuations, figures, and filler words in social media posts, which do not help to identify cyberbullying, but they produce noise. Therefore, the process of preprocessing is essential in order to improve the quality of input data to ML models. The preprocessing piping encompasses:

- Elimination of stop words: Narrowing of common words (e.g., words like and, the, etc.) which provide little semantic interest.
- Cutting out punctuation and numbers: Cutting characters which do not carry textual meaning.

One: The process of dividing sentences or words into tokens is called tokenization.

- Stemming/Lemmatization: Removing words to their root or base to standardize variations (ex: running to run).
- Textual data is cleaned and subsequently converted into numerical forms that are understandable by the ML algorithms. Two features extraction methods are prominent:
- Bag-of-Words (BoW): Turns the text into a vector of counts of occurrences of words, and taking each word as an equal weight.
- Term Frequency-Inverse Document Frequency (TF-IDF): Words are weighted by how prevalent they are within a document in comparison with how prevalent they are within the entire dataset, laying more weight on those words that are most informative about discrimination.

Machine Learning Classification Techniques

The processed feature vectors are fed into some machine learning classifiers to detect the content of cyberbullying. The available algorithms that are assessed in the study include:

1. Decision Tree (DT):

A full-fledged classification and regression model that recursively breaks down data on the basis of feature values. The internal nodes indicate a decision rule and the leaf nodes indicate a class label. The model walks through the tree and classifies each input case giving interpretable decision paths [23].

2. Naive Bayes (NB):

A Bayes-based probabilistic classifier in a case of independence between features. It estimates the posterior probability of classes given the features and classify inputs using the best probability. It is simple and efficient and is applied to binary and multi-class problems [24]. Formally, for a feature vector

X= (x1, x2,…,xn) and class y, NB estimates P(y|X).

3. Random Forest (RF):

A group of several decision trees which work by making each prediction and combining them through majority voting. This method increases accuracy and minimizes overfitting as compared to single trees and provides strong classification [25]. To illustrate, when the majority of trees predict class B the RF classifier will predict class B.

4. Support Vector Machine (SVM):

A strong classifier determining an optimum hyperplane between classes with a maximum margin in the feature space. SVM is capable of both linear and non-linear classification using the higher-dimensional spaces of data through the use of kernel functions. The linear kernel uses the dot product of the input vectors which is:

$K(xi, xj) = xi^{T}xj$

SVM typically provides high accuracy and computational efficiency, making it well-suited for cyberbullying detection tasks [26].

Experiment and Results

This paper presented and tested four machine learning models, i.e., Decision Tree (DT), Naive Bayes (NB), Support vector machine (SVM), and RF to analyze social media comments as either bullying or non-bullying. This part gives an insight into the datasets employed in the experimentation process, and it outlines the experimental setup as well as the results of the experiment.

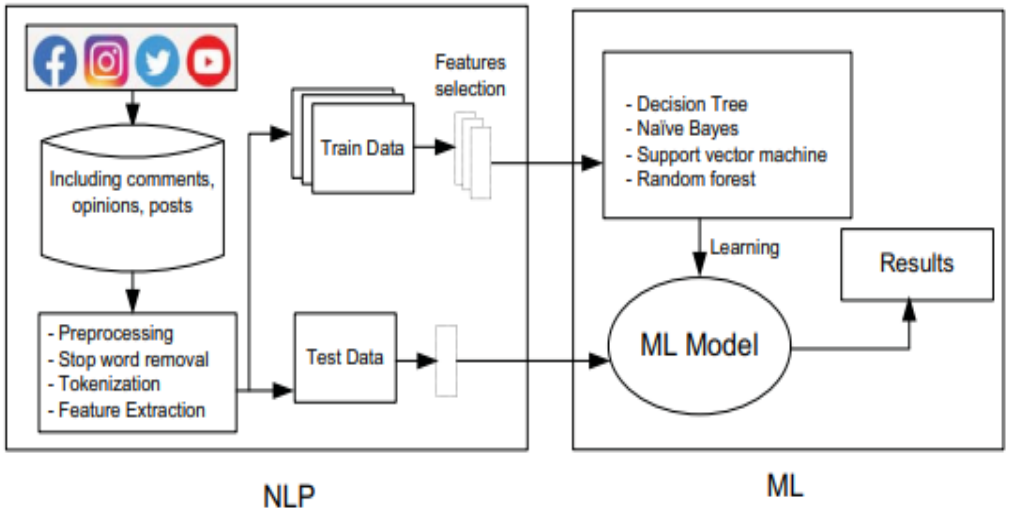
Datasets

We used two different datasets in our experiments:

- Dataset 1: This dataset consists of remarks of different Facebook posts and includes equal proportions of bullying and non-bullying remarks as a reflection of real-life social media communication.
- Dataset 2: This dataset comes out of Kaggle and is comprised of tweets that are either bullying or non-bullying. It is the style and content of a different social media platform, providing further variety to model testing.

All the entries in both datasets were manually annotated as being in one of two categories:

- Non bullying: These comments contain positive supportive, or neutral comments that show no bullying behaviour. An example to illustrate non threatening and constructive feedback is a comment like this photo is very beautiful.
- Bullying: Comments in this category contain offensive language, harassment, threats, or other forms of abusive communication indicative of cyberbullying.



Bullying Text Example and Evaluation Metrics

Social media language that is used by bullies or offenders to harass others will most definitely contain offensive words and expressions that are aggressive in nature. As an example, the saying go away, bitch is a widely known cyberbullying phrase because it is abusive and derogatory. In order to identify such language automatically, machine learning algorithms are coded in popular Python libraries such as scikit-learn, TensorFlow, or PyTorch.

The effectiveness of such algorithms is measured by metrics based on the confusion table (also known as a contingency table) such as Table I [28]. The four values constituting the confusion matrix compare the predicted labels to actual labels.

- True Positive (TP): Bullying instances which are correctly distinguished as bullying.
- False Positive (FP): The cases of non-bullying that are falsely recognized as bullying.
- False Negative (FN): The amount of bullying cases that are mistakenly defined as non-bullying.
- True Negative (TN): The number of non-bullying cases which are correctly identified as non-bullying.

The values are fundamental in the computation of significant evaluation metrics as accuracy, precision, recall, and F1-score. As an example, a large false-positive rate means that the model is incorrectly labeling many non-bullying remarks as bullying, whereas large false-negative rate implies that the model is not identifying bullying.

$$\text{Accuracy} = \frac{\sum \text{True Positive} + \sum \text{True Negative}}{\sum \text{Total Samples}}$$

Table I: The Confusion Matrix

	Condition Positive	Condition Negative
Predicted Condition Positive	True Positive	False Negative
Predicted Condition Negative	False Positive	True Negative

$$\text{Precision} = \frac{\sum \text{True Positive}}{\sum \text{Predicted Condition Positive}} \quad (5)$$

$$\text{Recall} = \frac{\sum \text{True Positive}}{\sum \text{Condition Positive}} \quad (6)$$

Receiver Operating Characteristic (ROC) Curve

To produce the ROC curve the true positive rate (sensitivity) acting on the false positive rate (1 - specificity) is plotted against the threshold of a diagnostic test [29]. This curve shows well the trade off between sensitivity and specificity: the more sensitive, the less specific. A more precise test results in a curve which is close to the upper-left corner of the ROC space, equating to high true positive and low false positive.

Dataset-1: Experimental Results

Dataset-1 comprises of user comments shared in different posts in Facebook. In order to assess the performance of various machine learning algorithms, we derived features in two methods, namely Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). The models were trained and tested on these feature representations.

Figure 2 and 3 show the accuracy and precision of the individual classifiers. The findings show clearly that Support Vector Machine (SVM) is more accurate and precise than other classifiers. Furthermore, TF-IDF feature trained models will always outperform BoW feature trained models. This is because TF-IDF can focus more on contextually relevant words and thus better the model can distinguish between bullying and non-bullying text.

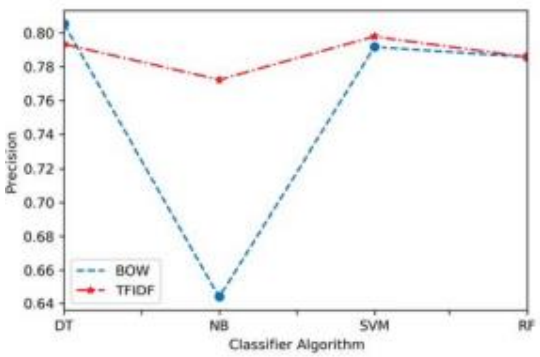


Fig. 2. Precision for Dataset-1

In Figures 4 and 5, we can see the ROC curves for the two different capabilities. To apply TF-IDF with BoW

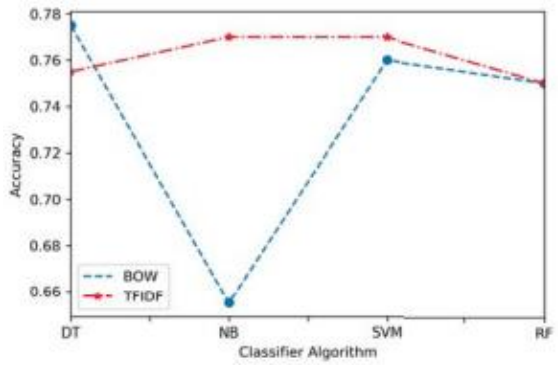


Fig. 3. Accuracy for Dataset-1

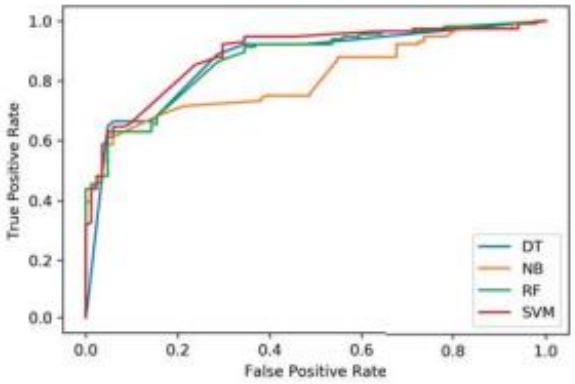


Fig. 4. ROC curve for BoW

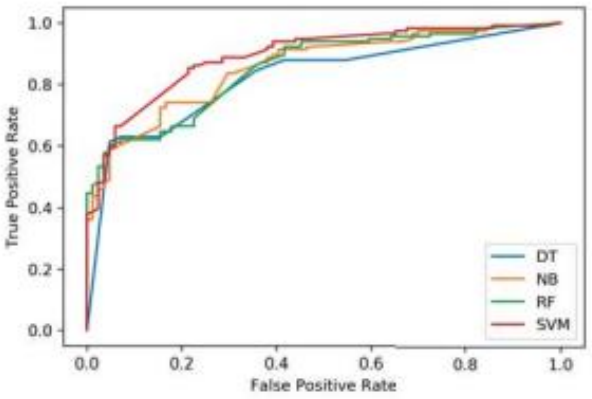


Fig. 5. ROC curve for TF-IDF

Results for Dataset-2

Accuracy and precision graphs of most machine learning methods are shown in figure 6 and figure 7. Regularly, we found out that TF-IDF is more accurate than BoW. SVM is superior to any other machine learning algorithm.

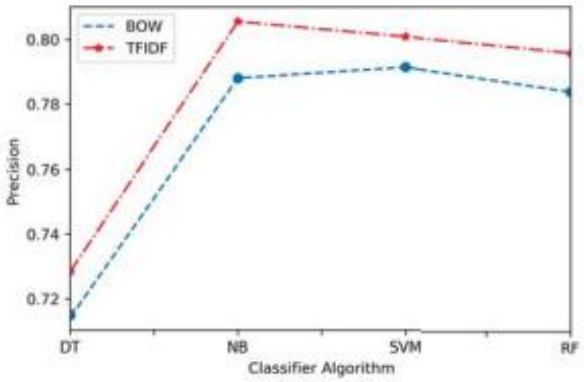


Fig. 6. Precision for Dataset-2

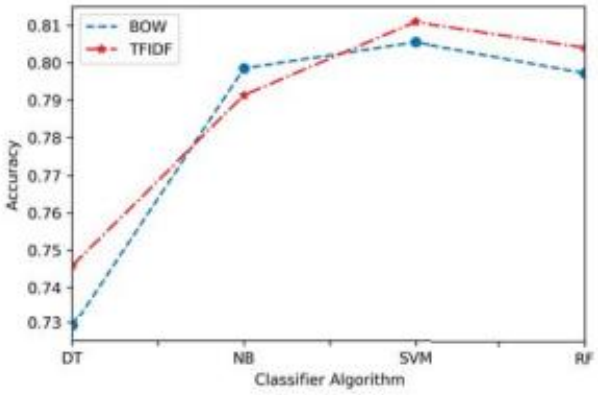


Fig. 7. Accuracy for Dataset-2

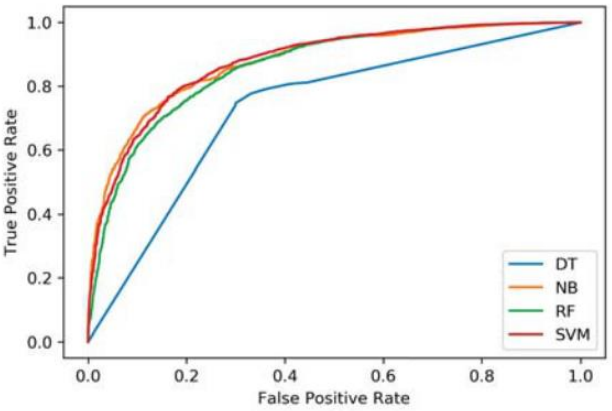


Fig. 8. ROC curve for BoW

Conclusion

The popularity of social media, especially among adolescents, has led to a significant rapid rise in cyberbullying, which has become a serious psychological, social, and legal problem. This pattern highlights the imminent necessity of smart and scalable and language-insensitive platforms that could identify and eliminate online abuse in real-time. This need

was met in this study where we investigated the automatic detection of posts in cyberbullying by examining two most popular text feature extraction methods; Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). We tested four machine learning classifiers including Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine (SVM) where SVM was proven to be the most successful in the two sets of features.

The implication of our findings is complex:

Technological Implication: The paper shows the effectiveness of classical machine learning models, especially SVM, in the early-stage detection of cyberbullying, which forms a computationally efficient basis of real-time content moderation tools.

Social Implication: The proposed solution can help make the social media and educational institutions a safer place by providing automated detecting tools, which will help decrease the risk of unsafe online use by the most vulnerable categories of users.

Academic and Research Implication: The findings confirm the importance of feature engineering and classic machine learning to cyberbullying studies. Furthermore, the present study will be used to establish a standard to be followed in a future comparative research on the deep learning and multilingual text classification.

Policy and Ethical Implication: Concrete cyberbullying detection tools can assist policy-makers and online platforms in creating ethical and transparent content regulation measures in line with digital rights and user protection legislations.

In the future, it is hoped that by incorporating deep learning architectures, including recurrent neural networks (RNNs) and transformers, into a single framework the adaptability of the model can be improved. Also, we are seeking to increase the reach of our system by adding multilingual datasets, and the preliminary goal is to identify and classify cyberbullying in Bengali text. This will greatly expand the range of automated cyberbullying prevention systems and offer more inclusive digital safety solutions to more varied language and cultures.

References

1. Fuchs, C. (2017). Social media: A critical introduction. Sage Publications. Selwyn, N. (2012). Social media in higher education. The Europa World of Learning, 1(3), 1–10.
2. Karjaluoto, H., Ulkuniemi, P., Keinänen, H., & Kuivalainen, O. (2015). Antecedents of social media B2B use in industrial marketing context: Customers' view. Journal of Business & Industrial Marketing.
3. Akram, W., & Kumar, R. (2017). A study on positive and negative effects of social media on society. International Journal of Computer Sciences and Engineering, 5(10), 351–354.
4. Tapscott, D., Ticoll, D., & Lowy, A. (2015). The digital economy. McGraw-Hill Education. Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., Desmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites: An

experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. Computers in Human Behavior, 31, 259–271.

5. Hoff, L., & Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. Journal of Educational Administration.
6. Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. Archives of Suicide Research, 14(3), 206–221.
7. Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on Web 2.0. Proceedings of the Content Analysis in the WEB, 2, 1–7.
8. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In Proceedings of the Social Mobile Web. Citeseer.
9. Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In 2011 10th International Conference on Machine Learning and Applications and Workshops (Vol. 2, pp. 241–244). IEEE.
10. Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710.
11. Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In European Conference on Information Retrieval (pp. 141–153). Springer.