Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

Cyberbullying Detection in Twitter Conduction Graph Mining and Machine Learning

Fatima N. Ali Hussein* Hiba J. Aleqabie*

*Collage of Computer Science and Information Technology /University of Kerbala.

Article Info

Article history:

Received 8-6-2023

Received in revised form 20-6-2023

Accepted 20-6-2023

Available online13-12 -2023

Keywords: cyberbullying detection, NLP, machine learning, graph mining, social media.

Abstract

Cyberbullying has become a severe problem as a result of the extensive use of social media platforms. It mainly refers to the act of utilizing digital methods to intentionally hurt, harass, or intimidate a person or group of people. Cyberbullying can take place on a variety of social media sites, including Facebook, Twitter, Instagram, and Snapchat. This form of bullying can have several negative impacts on individuals, including psychological distress, social isolation, academic problems, and even physical harm. In this study presents an approach for detecting cyberbullying on Twitter using graph mining and machine learning techniques. The study extract features of centrality measures in addition to the standard feature of texts TF.IDF and BOW. Machine learning algorithms are employed to train models for cyberbullying detection, supervised learning techniques. Where it trained using labeled data, Experiments on a large-scale dataset from Twitter were conducted to evaluate the effectiveness of the approach. The results showed that combining graph mining techniques with machine learning significantly improved the accuracy and efficiency of cyberbullying detection in Twitter. Specifically, when utilizing the Random Forest model with positive feature, we achieved a perfect accuracy rate of 0.98%.

Corresponding Author E-mail : Fatima.n@uokerbala.edu.iq , Hiba.jabbar@uokerbala.edu.iq

Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

Nowadays, internet technology is used nearly exclusively for communication, which might encourage negative or hazardous behaviors. Cyberbullying is a notable instance of such disruptive or harmful behavior. According to research, cyberbullying is characterized by a shift from offline to online bullying tactics via social media platforms[1]. Cyberbullying is described as an aggressive, intentional act that is committed by a group or an individual against a victims who is unable to defend themselves readily using electronic, digital, multi-modal modes of contact, messaging, and communication constantly[2]. Figure 1 depicts the most prevalent social Media network worldwide in 2022. Facebook had approximately 2,936 million users in 2022, making it the most prominent social networking site. There will be 7.98 billion people in the globe in 2022, 5.03 billion internet users, 5.34 billion mobile phone users, and 4.70 billion people who use social media regularly [3].

According to one study [4] , the disclosure of various types of cyberbullying increases suicide thoughts among youths. Despite measures, the reconstruction of cyberbullying victims is difficult for society and families. Self-hatred, dominance, isolation, and reactions to the socialization technique result in problematic and unhappy adults.

Furthermore, this mental imbalance may be enough to create future bullies. A text comment is a popular kind of cyberbullying. The most effective method for detecting cyberbullying and managing these social difficulties is to use a machine learning model. Furthermore, artificial intelligence (AI), namely NPL, can be utilized to prevent text-based bullying [5].

2. THEORETICAL BACKGROUND

The term "cyberbullying" refers to any form of bullying conducted via the Internet or other digital platforms. In the form of chats, texts, messages, comments, forum posts, and images, students may be cyberbullied on their phones, laptops, and other devices. Cyberbullying refers to any repeated communication of hostile or offensive remarks by a person or group on social media with the goal of causing damage or distress to others[6]. A branch of computer science called Natural Learning Processing (NLP) aims to make it easier for computers and people to communicate with one another. Its major goal is to develop an automated setting that can comprehend human language and utterance meanings. NLP is incredibly important since it significantly affects how we conduct our everyday lives[7].

Mutual information (MI) is another commonly used feature selection technique in machine learning. MI measures the mutual dependence between two variables, such as a feature and the target variable.

By using MI, machine learning algorithms can identify which features have the strongest dependence on the target variable, and use only those features for prediction. This can lead to greater accuracy and reduced overfitting compared to using all available features.

Additionally, MI can help identify and remove redundant features, which can reduce the dimensionality of the dataset and improve algorithm efficiency. This is particularly important for high-dimensional datasets, where feature selection can significantly improve prediction performance[8].

Using the $\chi 2$ statistic, machine learning models can identify important features and reduce the dimensionality of the dataset by selecting only the most relevant features. This can lead to more efficient and accurate models, as it reduces the risk of overfitting and improves the generalization of the model to new data. Therefore, the use of chi-square ($\chi 2$) in machine learning can have a significant positive impact [9].

Support Vector Machines (SVM) use the separation margin between data points of different classes as a classification criterion. The algorithm reduces the dimensionality of the original feature space as defined by the user. By optimizing the margin distance, support vectors are identified, which are the data points closest to the separating hyperplanes [10].

The Random Forest classifier comprises multiple decision tree classifiers [11]. Each tree independently predicts a class, and the final outcome is determined by selecting the class with the highest number of predictions. This classifier is a supervised learning model that delivers accurate results by merging the predictions of multiple decision trees.

Graph mining involves the process of extracting non-trivial graph structures from a single graph or a collection of graphs. It begins with feature extraction, where all text is transformed into a graph. The bag-of-words approach is a common technique used for this purpose, representing words in a text as a graph. To train the graph mining algorithm, labelled training data containing graphs derived from various text samples is used. This training process is employed to create a classification model[12]. The Centrality is a feature falls under the category of informative score features. The sentence's centrality implies that it is similar to other sentences. A document (or a collection of documents) is represented as a graph, with nodes representing sentences and connections connecting them weighted according to their similarity. The centrality of a node can be determined by computing its degree or by running a ranking algorithm. After calculating the centrality score for each sentence, the sentences are sorted in reverse order, with the highest-ranking ones included in the summary. If a sentence has a greater centrality degree, it is the best contender for inclusion in the summary, and its score is calculated as follows in (equation 1) [36].

$$\textbf{Centrality}(\textbf{Si}) = \sum_{i=0}^{n} \mathrm{CosSim}(\mathrm{Si}, \mathrm{S}_{(\mathrm{n}-\mathrm{i})}) \quad \text{.....(1)}$$

Where Si represent sentence and CosSim is mean cosine similarity distance.

3. LITERATURE REVIEW

As per the findings of [11], a study was conducted to identify rude and harassing messages in the English language on social media platforms, utilizing four distinct machine learning algorithms, namely SVM, RF, NB, and DT. The four machine learning techniques were evaluated for their accuracy using two distinct features, namely TF-IDF and Bow. The datasets pertaining to Facebook and Twitter were effectively extracted from the online platform Kaggle.com. The findings indicate that Support Vector Machines (SVM) outperformed all other machine learning techniques employed in the investigation. Similarly, TF-IDF surpassed Bow.. [12] The results of their trials on Multinomial Nave Bayes, SVM, and the k-nearest neighbor's method on Reuters R8 reveal that SVM performs better. While in [13] The Bag of Word (BOW) method is used to weight each word, which uses three weighting features (unigram, bigram, and trigram), while outcomes show that for the measurement of accuracy weighting based on features and algorithms, the SVM classification algorithm exceeded other algorithms by 76%. By 76%, the Decision Tree classification algorithm outscored the other techniques.

In a recent research conducted by [14], the impact of various feature selection techniques, such as MI (mutual information) and Chi2 (chi-square), on the accuracy of different machine learning models was examined. The study revealed that both MI and Chi2 had a beneficial effect on the classification accuracy. Furthermore, when these two feature selection techniques were combined, the highest accuracy was achieved.

An experiment was conducted by researchers to implement Support Vector Machine (SVM) for tweet classification, utilizing four distinct kernels, namely Polynomial, Radial Basis Function (RBF), Sigmoid, and Linear. The SVM technique has been identified as an effective way of categorizing instances of cyberbullying, as per the results of various research investigations. Additionally, it is known that the sigmoid kernel has the highest level of accuracy, reaching 83.85%. [15]. On the other hand in [16] Random Forests, Recurrent Neural Network (RNN), SVM, POS tags, function words, and content word features Bidirectional Long Short-Term

Memory (Bi-LSTM), and Multinomial Naive Bayes (MNB) (RF) were used to achieve accuracy of 90.45% based on data from Twitter. While in [11], utilized two different features, Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF), and applied four machine learning techniques, including Support Vector Machines (SVM), to recognize bullying content. The evaluation revealed that the Naive Bayes (NB) classifier achieved an accuracy rate of approximately 80% using BOW and 79% using TF-IDF. Furthermore, the SVM classifier performed slightly better, with accuracies of around 80% for BOW and 81% for TF-IDF.

4. METHODOLOGY

As can be seen in Figure 1, the work that is planned will involve a number of stages, which will now be taken down as follows:

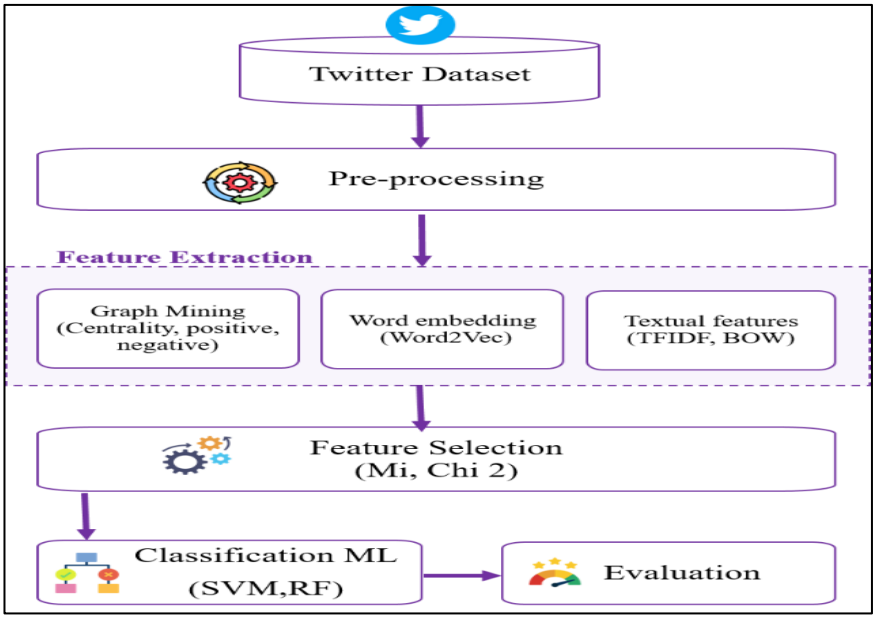

Figure1. the proposed work

4.1 Twitter Dataset

The dataset used in this research to detect cyberbullying was collected from the Kaggle website, specifically from www.kaggle.com. The dataset contains approximately 47,000 entries, but for this study, only 6,000 entries were used. The tweets in the dataset were gathered during the COVID-19 period when social media played a significant role in connecting people globally. According to statistics from UNICEF, during that time, 87% of individuals observed cyberbullying, and 36.5% of people actually experienced it. The dataset was split into 80% for training purposes and 20% for testing [3].

4.2 Pre-processing

This stage consists of several steps include:

1. Tokenization: The document is tokenized into words
2. Removing URLs: Any words that match the pattern of a URL (starting with 'http' and followed by non-whitespace characters) are removed from the list of words.
3. Removing non-alphabetic characters: Any non-alphabetic characters are removed from the words using regular expressions (re.sub('[^a-zA-Z]', '', word)).
4. Lowercasing: The words are converted to lowercase.
5. Removing Stopwords: removing customized Stopwords (common words like "the," "and," "is," etc.) that may not affect the accuracy.
6. part-of-speech tagging is performed on the tweets. The resulting tagged words are stored in the tagged words variable as a list of tuples, where each tuple contains a word and its corresponding part-of-speech tag. Next, a list comprehension is used to filter the tagged words and select only the nouns, assuming bullying terms are nouns.

4.3 Feature Extraction

Following pre-processing, the extraction procedure is used to identify qualities that significantly affect a phrase and decrease the dimensions of the data. Word weighting is the term for this method[13].

The specified model converts data into a format suitable for applying machine learning algorithms. To identify the most important features in the dataset, the TF-IDF vectorizer and Bag of Words (BOW) algorithms are commonly used. These algorithms extract and organize the significant features as a list. TF-IDF is a widely used method in text mining that assesses the value of words through statistical analysis in a collection of documents [7].The Word2vec model takes into consideration both syntactic and semantic similarities between words.

TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

The weight is typically calculated using the normalized Term Frequency (TF). The computation involves the division of the frequency of a given term (t) within a specific document (d) by the overall count of words present in that document. The Inverse Document Frequency (IDF) is computed by taking the logarithm of the ratio of the total number of documents to the sum of the number of documents which include a given term (t). The computation of the TF-IDF weightfor each term in the corpus involves the multiplication of the term's term frequency (TF) by the inverse document frequency (IDF). equation 2 shows the TF.IDF.

$$TF - IDF(wt) = TF_{t,d} \times \log\left(\frac{N}{df_t}\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

BOW, or Bag-of-Words, is a method of representing text data by creating fixed-length vectors. It involves counting the occurrences of each word in the text. When applied to tweets, each tweet is treated as input data, and the frequency of each word or phrase in the tweet is determined. This process generates a vector-based representation based on the numerical values of the terms [7].

The following types of features were used:

1. Word2Vec embeddings pre-trained to create tweet-level feature representations. The embeddings used were 400-dimensional and were sourced from a dataset of 10 million tweets from the Edinburgh corpus [17].
2. Sentiment feature vector was obtained by using SentiStrength [18]. It calculated the positive and negative scores for each tweet, providing a representation of the sentiment expressed in the tweet.

5. Feature selection

Feature selection techniques such as mutual information (MI) and chi-squared (Chi2) test play an important role in improving the classification accuracy of machine learning models.

The feature selecting and selection step involves manipulating and choosing relevant variables or features from the available data. It includes techniques such as transforming variables, creating new features, and selecting the most informative ones for a particular task.

5.1. Mutual Information

Mutual information (MI) is a commonly used feature selection method in machine learning and has been studied extensively in recent years. The corpus contained mutual information values for each word, indicating the statistical relationship between words in terms of their information content [17] as seen in equation (3).

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y) \quad \text{.....(3)}$$

5.2 Chi_2

the effectiveness of this techniques may vary depending on the specific dataset and machine learning models being used. Therefore, it is important to carefully evaluate the impact of these techniques in each specific case. as seen in equation (4).

$$\chi 2 = \sum(Oi - Ei)2/Ei \quad \text{.....(4)}$$

5. Result

The performance of SVM and Random Forest models was assessed to determine which one was better at detecting cyberbullying in a Twitter dataset. Accuracy, precision, recall, and F1 score metrics were used to evaluate the models' effectiveness. Figure 1 presents the overall performance of the two models based on these metrics. Upon analyzing the results, it was observed that both models performed comparably in detecting cyberbullying on social media.

The SVM model using combination graph features and word embedding (degree centrality and word2Vec) achieved an accuracy of 0.87. And get 0.98 when utilizing textual feature TFIDF and feature selection Mutual information (MI) with graph features (positive feature), and 0.97 with the negative feature. On the other hand, the Random Forest model with degree centrality obtained an accuracy of 0.93, and 1.00 and 0.99 with the positive and negative features, respectively. When utilizing the SVM model with textual features, the accuracy was 0.95, 0.95, and 0.70 with TF-IDF, BOW, and Word2Vec, respectively. Similarly, the Random Forest model obtained an accuracy of 0.97, 0.96, and 0.88 with TF-IDF, BOW, and Word2Vec, respectively. Comparing the results, the Random Forest model with positive graph feature yielded the best performance.

Table1. the machine learning result with different features

Algorithm	Graph Features	Accuracy	Text Features	Accuracy
SVM	Degree Centrality	0.87	TF-IDF	0.95
	Positive Features	0.98	BOW	0.95
	Negative features	0.97	Word2vec	0.64
RF	Degree Centrality	0.93	TF-IDF	97%
	Positive Features	0.98	BOW	96%
	Negative features	0.99	Word2vec	88%

6.DISCUSSION AND CONCLUSION

The utilization of the internet and social media platforms brings about evident benefits to communities; however, their widespread usage can also lead to substantial negative outcomes. These include instances of unwanted sexual exposure, cybercrime, and cyberbullying. We develop a model for detecting cyberbullying with twitter platform. In this study, we aim to investigate the effect of different graph-based features, specifically Centrality, Closeness, and their combination, on the performance of SVM and RF algorithms using three feature representation methods, namely, graph-based features (GBF), term frequency-inverse document frequency (TFIDF), and word2vec.

Our experiments show that the SVM model performs better when using the GBF feature representation method compared to the other two methods. The RF algorithm's superior performance using the TFIDF feature representation approach may be attributed to its enhanced ability to extract patterns and textual information from the input. In natural language processing, TFIDF (Term Frequency-Inverse text Frequency) is a widely used approach that gives terms weights depending on how frequently they appear in a text and how seldom they occur in all documents. This makes it possible for the algorithm to concentrate on key words and standout phrases, which is especially helpful for text-based data. Conversely, graph-based features use concepts from graph theory, including nodes and edges, to describe interactions between data points. These characteristics have the ability to capture intricate connections and dependencies between data pieces, which might be useful for particular kinds of information or issues. However, it appears that in this particular instance, the RF algorithm is not as successful in making use of these graph-based properties as it is with TFIDF. It's important to remember that these answers are conjectural and would need to be verified by more research and analysis. The particular dataset, issue, and algorithm being employed can all affect how well a certain feature representation technique performs.

Furthermore, our research found that combining textual and Graph information significantly improves the performance of SVM and RF models. We can capture both local and global structural information inside the graph by combining these two types of characteristics. This complete technique is a powerful tool for reliable data categorization.

Comparing the results of our proposed system with several research studies, it is evident that our approach offers numerous significant benefits over previous machine learning with graph mining systems.

Firstly, our approach leverages cutting-edge graph mining algorithms, which empower us to extract valuable insights and patterns from intricately linked data. By utilizing these advanced algorithms, we are able to uncover hidden dependencies and correlations that would remain unnoticed when using more conventional machine learning techniques. This is achieved by effectively utilizing the inherent links and structures present in the data.

Furthermore, our algorithm surpasses the limitations of traditional machine learning with graph mining by incorporating scalable methods and integrating domain-specific information. This integration allows us to efficiently extract important insights from complex interrelated data, surpassing other methods currently employed in this industry.

To demonstrate the effectiveness of our approach, we have presented the results in Table 2. The results clearly showcase how our system outperforms existing methods in terms of extracting meaningful insights from complicated interrelated data.

Table 2 The Compression Results

Ref.	Features	Techniques	Performance (accuracy)	Dataset
[7], 2022	Bag Of Word (BOW), unigram, bigram, and trigram	SVM, DT	76%	Twitter.
[8], 2022	Polynomial, Radial Basis Function (RBF), Sigmoid, and Linear.	SVM	83.85%	Twitter.
[10], 2021	BOW, TFIDF	DT, RF, SVM, NB	80%	Facebook
[11], 2023	BOW, TFIDF	SVM	95%	YouTube, Twitter
[3], 2022	TFIDF	LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost	85.5%	Twitter
Proposed work	TFIDF, BOW, Word2Vec	SVM, RF	95%	Twitter

In conclusion , the proposed system stands out due to its utilization of cutting-edge graph mining algorithms, ability to reveal hidden dependencies and correlations, scalability, integration of domain-specific information, and superior efficiency in extracting important insights from complex data. It also, highlights the significance of feature selection in machine learning algorithms, specifically when utilizing graph-based features. The incorporation of Degree Centrality and Word Embedding features exhibits encouraging outcomes in enhancing the accuracy of SVM and RF models. These findings hold potential for application in diverse domains such as natural language processing and social network analysis.

References

[1] G.M. Abaido, Cyberbullying on social media platforms among university students in the United Arab Emirates, Int. J. Adolesc. Youth. 25 (2020) 407–420. https://doi.org/10.1080/02673843.2019.1669059.

[2] A. Ibtihaj, J. Alves-Foss, Cyber_Bullying_and_Machine_Learning_A_Su.pdf, (n.d.). https://doi.org/https://doi.org/10.5281/zenodo.4249340.

[3] M.I. Mahmud, M. Mamun, A. Abdelgawad, A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning, 2022 IEEE Glob. Conf. Artif. Intell. Internet Things, GCAIoT 2022. (2022) 166–170. https://doi.org/10.1109/GCAIoT57150.2022.10019058.

[4] S. Neelakandan, M. Sridevi, S. Chandrasekaran, K. Murugeswari, A.K. Singh Pundir, R. Sridevi, T.B. Lingaiah, Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media, Comput. Intell. Neurosci. 2022 (2022). https://doi.org/10.1155/2022/2163458.

[5] M. Mamun, A. Farjana, M. Al Mamun, M.S. Ahammed, Lung cancer prediction model using ensemble learning techniques and a systematic review analysis, in: 2022 IEEE World AI IoT Congr., IEEE, 2022: pp. 187–193. https://doi.org/10.1109/AIIoT54504.2022.9817326.

[6] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, A.R. Rajan, Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification, Comput. Electr. Eng. 92 (2021) 107186. https://doi.org/10.1016/j.compeleceng.2021.107186.

[7] A.M. Alduailaj, A. Belghith, Detecting Arabic Cyberbullying Tweets Using Machine Learning, Mach. Learn. Knowl. Extr. 5 (2023) 29–42. https://doi.org/10.3390/make5010003.

[8] H. Zhou, X. Wang, R. Zhu, Feature selection based on mutual information with correlation coefficient, Appl. Intell. (2022) 1–18.

[9] A. Sikri, N.P. Singh, S. Dalal, INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Chi-Square Method of Feature Selection : Impact of Pre- Processing of Data, 11 (2023) 241–248.

[10] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, M. Prasad, Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques, Electron. 10 (2021). https://doi.org/10.3390/electronics10222810.

[11] M.M. Islam, M.A. Uddin, L. Islam, A. Akter, S. Sharmin, U.K. Acharjee, Cyberbullying Detection on Social Networks Using Machine Learning Approaches, in: 2020 IEEE Asia-Pacific Conf. Comput. Sci. Data Eng., IEEE, 2020: pp. 1–6. https://doi.org/10.1109/CSDE50874.2020.9411601.

[12] H. Abdulla, W. Awad, Text Classification of English News Articles using Graph Mining Techniques, (2022) 926–937. https://doi.org/10.5220/0010954600003116.

[13] A. Muzakir, H. Syaputra, F. Panjaitan, A Comparative Analysis of Classification Algorithms for Cyberbullying Crime Detection: An Experimental Study of Twitter Social Media in Indonesia, Sci. J. Informatics. 9 (2022) 133–138. https://doi.org/10.15294/sji.v9i2.35149.

[14] A. Suruliandi, G. Mariammal, S.P. Raja, Crop prediction based on soil and environmental characteristics using feature selection techniques, Math. Comput. Model. Dyn. Syst. 27 (2021) 117–140.

[15] S.S. Wijayanti, E. Utami, A. Yaqin, Comparison of Kernels on Support

Vector Machine (SVM) Methods for Analysis of Cyberbullying, in: 2022 6th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng., IEEE, 2022: pp. 104–108. https://doi.org/10.1109/ICITISEE57756.2022.10057761.

[16] B.A.H. Murshed, J. Abawajy, S. Mallappa, M.A.N. Saif, H.D.E. Al-Ariki, DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform, IEEE Access. 10 (2022) 25857–25871. https://doi.org/10.1109/ACCESS.2022.3153675.

[17] B.A. Talpur, D. O'Sullivan, Cyberbullying severity detection: A machine learning approach, PLoS One. 15 (2020) e0240924. https://doi.org/10.1371/journal.pone.0240924.

[18] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Am. Soc. Inf. Sci. Technol. 63 (2012) 163–173. https://doi.org/10.1002/asi.21662.