

# Detection of Cyberbullying using Bullying Features and Machine Learning: A comparative study

Yashashvi Singh<sup>1</sup>, Bam Bahadur Sinha<sup>1</sup>[0000–0002–7284–9850], and Mohammad Ahsan<sup>2</sup>[0000–0002–2619–6529]

<sup>1</sup> Data Science and Intelligent Systems,  
Indian Institute of Information Technology Dharwad, Karnataka, India

<sup>2</sup> Computer Science and Engineering Department,  
National Institute of Technology Hamirpur, Himachal Pradesh, India  
bahadurbam43@gmail.com

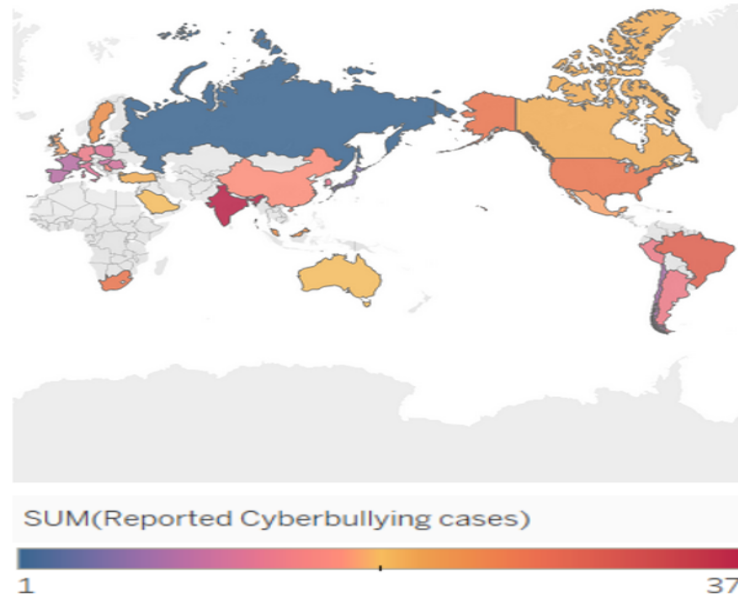
**Abstract.** Cyberbullying is a curse that has come upon humanity with the advent of digitalization. A large section of the population faces deteriorating comments about their bodies and personalities from strangers every day via social media applications. Teenagers face bullying at an escalating rate with the increasing popularity and accessibility of social media platforms. Cyberbullying not only abuses young minds and their emotions but also snatches away their prestige and confidence, leaving a permanent mark for the rest of their lives. In some cases, it may leave victims depressed and prone to suicide attempts. This problem is not new and has been the prominent reason for depression among teens for many years, but still, inconsequential steps have been taken to eradicate it. This paper aims to not only raise awareness about this social issue but also use different machine learning techniques, namely: the Naive Bayes approach, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM) for tackling cyberbullying in real-time. The efficacy of machine learning approaches for detecting cyberbullying is demonstrated via different performance measures such as precision, recall, F1-score, accuracy, and ease of implementation with respect to time.

**Keywords:** Social networking threats · Cyber Crime · KNN · SVM · Naive Bayes

## 1 Introduction

Bullying is repetitive, undesired, and aggressive behaviour that a victim suffers because of mean, hurtful, and damaging comments passed by a bully [1]. Bullying is more likely in situations where there is an unequal distribution of power among people. This power can be due to the bully having more physical strength, popularity, or knowledge of demeaning and embarrassing information about the victim. Bullying has always been persistent in society, but as we have now moved to the computer era, where we can utilise the services of connecting with friends and family remotely via social media, bullying has transformed

into a much more dangerous and vandalising form, cyberbullying. Social media platforms provide a way to post photos and views to create a permanent public record of one's profile. This profile serves as a digital image of the user. When negative comments are written on these platforms, they torment the digital reputation of the user. These comments can be seen by the other users as well, which further increases the impact of cyberbullying on the victim's minds [2]. Also, social media provides bullies with anonymity, giving them an upper hand in the power play against their victims. With the use of the internet, cyberbullying is not restricted to a particular place like school or college but can be used anywhere 24/7 to victimize. Figure 1 illustrates the cyberbullying cases reported globally in year 2018<sup>1</sup>. Democracy and participation were brought by the digital landscape, mainly social media, which allowed individuals to freely exchange and consume information. On the other hand, the anonymity of the internet has made it easier for criminals to target innocent individuals. Freedom of expression, combined with an unregulated internet world, has exacerbated the problem of cyberbullying. Figure 1 illustrates the same.



**Fig. 1.** Cyberbullying cases reported globally - 2018

With the adversity that cyberbullying can bring upon a user, there is an alarming need to find smart solutions and means to detect and prevent cyberbullying. This paper proposes the use of various machine learning based algorithms

<sup>1</sup><https://public.tableau.com/views/GlobalViewsonCyberbulling2011-2018/GlobalViewsonCyberbulling2011to2018?:showVizHome=no>

that are suited for this job. Machine learning is a set of computer algorithms that can learn themselves through exemplary experiences. This experience can be provided to the computer in the form of labelled data sets. Natural Language Processing via machine learning models can easily analyse patterns in text and learn to classify them. As our solution requires a parallel approach to identify hate content in comments and text messages, machine learning algorithms prove optimal. Also, machine learning is a cheaper alternative as compared to more advanced techniques incorporating deep-learning and neural networks. This makes it easy to implement on any public internet messaging platform. The paper is focused on implementing the bag of words algorithm using SVM (Support Virtual Machine) classifiers, Naive Bayes classifiers, and KNN (K-Nearest Neighbour) classifiers. These techniques are implemented on the "Suspicious Communication on Social Platforms" dataset from Kaggle<sup>2</sup>. They collected the data from Facebook and Twitter groups. The data includes abusive language, discriminating comments based on racism, color, or gender, and threats. Data tagging is done using two labels: non-suspicious content (0) and suspicious content (1). Observations on algorithms were made repeatedly by varying training and testing dataset split percentage and results are noted according to the algorithm's average performance.

The main contribution of this paper is to perform a comparative study of various machine algorithms that can be used as a solution and testing their efficacy based on the obtained precision, accuracy, implementation ease, F1-scores, and recall-scores, enabling readers to choose the algorithm that best suits their needs. This paper is organised as follows: Section II shows various related backgrounds in cyberbullying. Then, section III describes the experimental methodology. Section IV explains the proposed architecture. Section V showcases the experimental results and observations. Finally, section VI concludes and discusses future work in the domain.

## 2 Related Work

While cyberbullying is defined by the three factors of bullying: (1) purposeful action to hurt individuals, (2) recurring activity over time, and (3) unique inequity among both sides engaging in such behaviors, features particular to bullying should be added to the definition: Both the 24-hour activity and the anonymity of its users have helped to boost its audience. In the last few years, several research work has been done to create awareness about cyberbullying and to propose and develop machine learning and deep learning models[3] for the detection of cyberbullying on social media platforms like YouTube, Twitter, Facebook, and others [4] [5]. In an effort to unravel the effects of cyberbullying on "Female Emirati University Students", [6] presented a detailed study on cyberbullying culture in the Emirates. [7] explained how cyberbullying is

<sup>2</sup><https://www.kaggle.com/syedabbasraza/suspicious-communication-on-social-platforms>

different and far more dangerous than trivial forms of bullying. They further stressed on how cyberbullying is a major threat to the adolescent population on the internet as it can leave them feeling depressed, alone, hopeless, and lower their self-esteem. Though the study was unable to show any direct relationship between cyberbullying and suicide, respondents did confess to having suicidal thoughts after being victimised by cyberbullying. Another significant contribution was provided by [8] where the study aimed to show why young users are resistant towards reporting cyberbullying cases. [8] shows why one cannot rely only on laws preventing cyberbullying alone, as a large number of cyberbullying instances are not reported and tackled. Hence, there is an urge to deploy algorithms online that can prevent cyberbullying on the fly. Using Naive Bayes classifiers on datasets from Kongregate, Slashdot, and MySpace, [9] obtained an accuracy of 95.79%. [10] collected their data from four Taiwanese social sites. They used data mining techniques and showcased that sentiment is one of the important factors in identifying cyberbullying as it allows us to understand the intentions of the user. Their work has a precision of 79%. Another approach [11] where they used feature extraction methods like bag-of-words, N-gram, and TFIDF to create an input vector. SVM, logistic regression, and random forest models were used by them. Logistic regression and random forest models performed equivalently while classifying the data. After the training and testing phase, the random forest was discovered to be the most accurate model out of the three, with an accuracy of 93%. While logistic regression and SVM models were accurate to 86% and 72% of the time, respectively.

An app "BullyBlocker" [12] which works on top of a machine learning model detects cyberbullying and sends appropriate alerts to parents if any bullying activities are detected within 60 days. Similar [13] proposed an approach where they implemented the SVM and Naive Bayes Classifiers on a Kaggle dataset. They found that the Naive Bayes classifier had an accuracy rate of 92.81% and the SVM classifier had an accuracy rate of 97.11%. [14] adopted different NLP techniques to identify cyberbullying using Sentiment Analysis. They classified their data binarily (positive/negative) based on the presence of hateful content. The dataset was labelled manually and a linear SVM was trained which gave 89% accuracy. Their work depicts that common classifiers and features can detect cyberbullying successfully. [15] continued cyberbullying detection using sentimental analysis, but they also introduced a new feature, emoticon icons. They tested their approach using J48, Naive Bayes, and SVM algorithms. The SVM algorithm triumphed among the other algorithms with an accuracy of 81%. [16] proposed a framework known as SICD where they used KNN classifiers. KNN classifiers achieved a 0.6105 F1-score and 0.7539 AUC score using this approach.

Most of these studies were based on supervised machine learning and used readily available classifiers like Naive Bayes, Support Vector Machine, K-Nearest Neighbor, and logistic regression to detect cyberbullying content across social media platforms. Also, one can observe that the dataset used in these studies was either taken from Kaggle or was collected and labelled manually from popular

social media sites like Facebook, Twitter, YouTube or from gaming sites that have a messaging service for the players. NLP has been extensively used as cyberbullying detection is closely linked to text analysis. This research paper aims to incorporate factors from these studies and explore and compare three machine learning techniques, out of which Support Vector Machine and Naive Bayes are widely used and K-Nearest Neighbor is comparatively less worked upon in this domain.

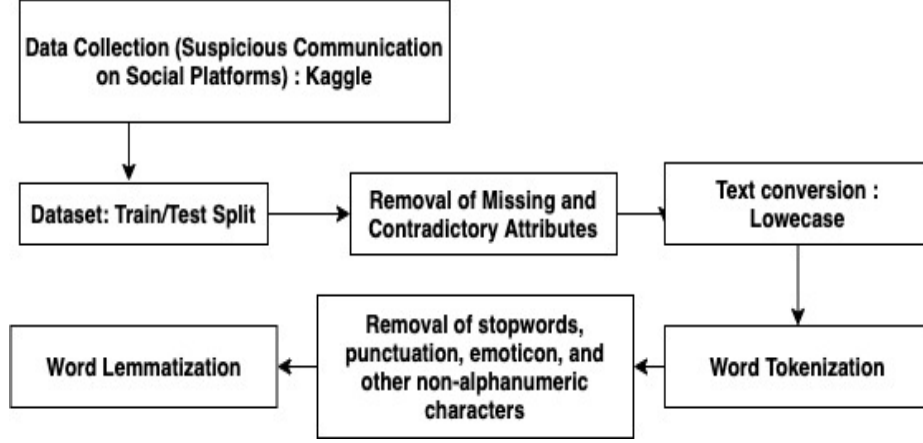
### 3 Experimental Methodology

This paper is a comparative study of three machine learning approaches used for the detection of cyberbullying on social media platforms, which are SVM, Naive Bayes, and KNN classification. These classifiers are pre-trained and are used on the "Suspicious Communication on Social Platforms" dataset obtained from Kaggle. The dataset was collected from Facebook and Twitter groups and was binarily classified using "0" for a negative instance and "1" for a positive instance of cyberbullying. We chose this dataset as it contained a high amount of racist, abusive, and discriminatory content. Also, we made sure that the dataset is taken from popular social media platforms as they host a large number of users which can provide a variety of cyberbullying text patterns to learn from. This dataset was then split into two sub-datasets required for training and testing. This was done for three different cases where the testing to training sub dataset size ratio varied as 9:1, 8:2, and 7:3 respectively.

For each case, the data was cleaned of any mislabeled or contradictory attributes. Then, the data was pre-processed by changing all the text to lowercase and performing word tokenization, i.e., the sentences or phrases were converted into an array of words, punctuation, and emoticon symbols as its elements. From this array, stop words and non-alphanumeric characters were removed. This was followed up by word lemmatization. This is the basic process of the bag-of-words algorithm. The above process can be summarised in a flow chart illustrated by Figure 2. Then, the various classifiers (Naive Bayes, SVM, and KNN) were used to make the classifications. A timer was started using the time library to note down the time taken by each machine learning model to learn from the training data and classify the test data.

Finally, results were obtained by running the code several times and the average scores were noted. Below are the metrics that were used for comparing and evaluating the performance of the various machine learning techniques. These metrics were also imported from the sklearn.metrics library:

- The efficiency and ease of implementation were marked by the resources needed by the models, i.e., size of the training dataset and the time required in each case. The algorithm that performed better with the least resources was declared more efficient than others.



**Fig. 2.** Steps involved in data cleansing and pre-processing

- The accuracy of the algorithm was measured in terms of the number of correctly classified messages by using Equation 1.

$$Accuracy = \frac{TPD + TND}{TD} \quad (1)$$

- The recall[17] score calculates how many positive instances i.e. messages containing cyberbullying, are detected out of all positive instances. Mathematical representation is give by Equation 2.

$$Rec = \frac{TPD}{TPD + FND} \quad (2)$$

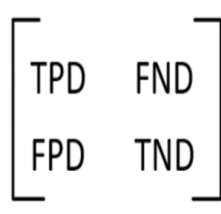
- Precision [17] given by Equation 3 measures the number of messages that actually contained signs of cyberbullying as compared to the number of messages reported by the model.

$$Pre = \frac{TPD}{TPD + FPD} \quad (3)$$

- F1-Score [17] is a metric computed using the harmonic mean of recall score and precision. F1-Score can be calculated using Equation 4

$$F1 - Score = 2 * \frac{Pre * Rec}{Pre + Rec} \quad (4)$$

- Confusion Matrix (CM) is a metric that represents the instances when the model gets confused while predicting the result. In the confusion matrix of a good model, diagonal elements are high and other elements are low. The confusion matrix representation for a model is shown in Figure 3. where, TPD = True Positive Datapoints



TPD	FND
FPD	TND

**Fig. 3.** Confusion matrix representation for a model

TND = True Negative Datapoints

FPD = False Positive Datapoints

FND = False Negative Datapoints

TD = Total Datapoints

As discussed earlier, the rationale for this paper is to provide a comparative study between commonly used ML classifiers in respect to cyberbullying detection.

## 4 Proposed Architecture

The proposed model can be designed by following 8 steps, as illustrated by Figure 4. The different steps involved in the proposed architecture are: Data collection, Train-Test splitting, Data cleaning, Data pre-processing (Bag of Words approach), Creating feature vector, Model training, Time computation for training the model, and Performance testing.

We tentatively explored our various options and chose a dataset that was well-suited based on the source of the data, size of the dataset, and whether it contained controversial topics like racism, sexist comments, etc. Since the data we obtained was raw, it contained many stop words, punctuation, emoticon icons, and user names. One example of sample text is represented by Figure 5.

The sample text message shown in Figure 5 is from the dataset that we obtained from Kaggle. This message has punctuation, stop words, and user names that need to be removed. The dataset also had missing and contradictory data points that were removed. The text was then converted to lowercase and tokenized. Then the removal of unnecessary stop words, punctuation, and other symbols was done. Finally, the text was lemmatized. This approach is called the Bag of Words approach. The Bag of Words takes in text data as input and breaks it into a set of words using the process of tokenization. Unique tokens are identified and the frequency for each data point is calculated. In addition, to frequency, the proposed model also takes into account the density of the tokens, i.e., the frequency of a particular token per total number of tokens, as suggested in [18]. After pre-processing and applying the bag-of-words approach, the sample text datapoint was transformed, and the result is illustrated via Figure 6.

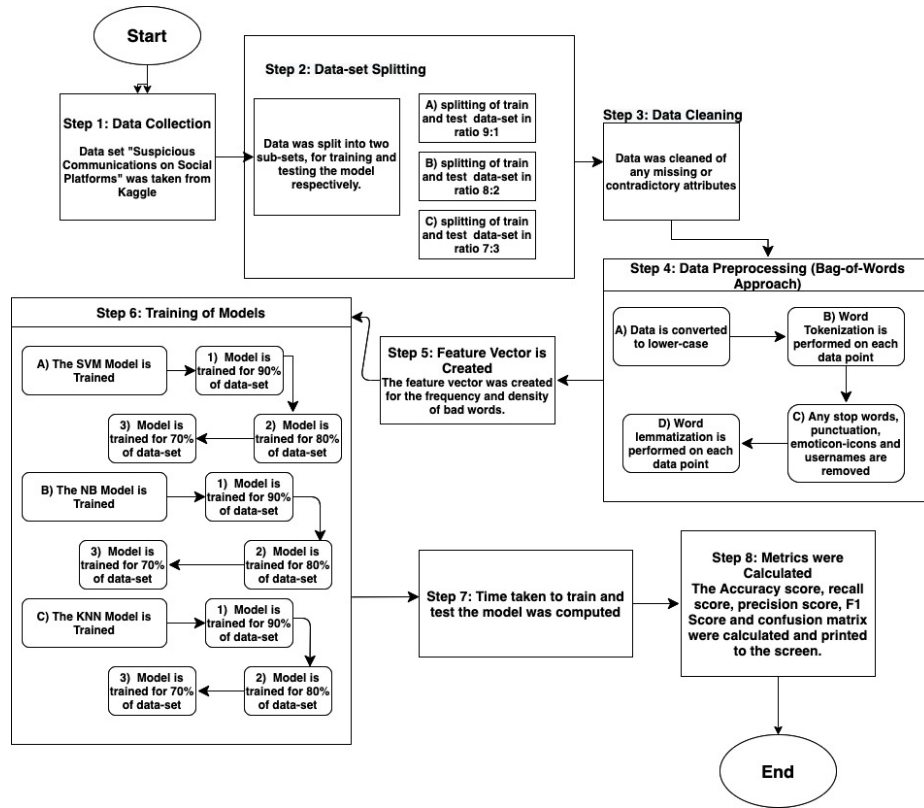


Fig. 4. Proposed architecture for detecting cyberbullying

& @kmellon: I'm 500 pages in AND at the day job. Eat that dick.	1
--	---

Fig. 5. Sample text from the dataset

['500','pages','day','job','eat', dick']	1
---	---

Fig. 6. Pre-processed sample text using Bag-of-Words



The dataset was divided into two subgroups. This was done for three different cases, with the test size for each being 10%, 20%, and 30%. Once the feature vector was made, a timer was started and a model was created using each classifier for all three cases. The model was fitted using the training dataset and finally, predictions were made using the testing set. The timer was stopped and the time taken was displayed on the screen. The accuracy score, F1 score, precision score, recall score, and confusion matrix were also recorded. In total, there were 9 different model configurations. The detailed results for each model configuration are discussed in the upcoming section.

## 5 Experimental Results and Observations

As already discussed, we obtained our data from Kaggle and conducted experiments using the bag-of-words approach combined with SVM, Naive Bayes, and KNN classification models. This was done for three cases where we varied the training dataset size available to the models. The results of the experiment are shown in Table 1. These results showcase that SVM outperforms other models in terms of accuracy as it has the highest accuracy for all 3 cases. Also, its F1 scores are the highest for all 3 cases, indicating a good balance between precision and recall for the model. The confusion matrix also indicates that the SVM model works pretty well for the detection of cyberbullying. SVM fails in its efficiency and implementation ease as it uses a large amount of time to make predictions. The time required for SVM is significantly higher than the other models. The KNN model outperforms the SVM model here, as it provides fairly close accuracy to the SVM but in exponentially less time. The KNN model's recall score and precision are also fairly close to the SVM model. The Naive Bayes (NB) model under-performs the most due to its low accuracy and precision. This model doesn't have a stable F1 score, indicating that it trades a high recall score for low precision. The Naive Bayes model also takes much less time than the SVM model.

## 6 Conclusion and Future Work

This paper presents a comparative study of three simple and pre-trained machine learning algorithms, which are Support Vector Machine (SVM) classifiers, Naive Bayes (NB) classifiers, and K-Nearest Neighbour (KNN) classifiers, on a common dataset obtained from Kaggle. Results of the experiments conducted help conclude that the K-Nearest Neighbour (KNN) algorithm proves to be the most suited algorithm among the three for the detection of cyberbullying in text messages shared via social media applications. The KNN algorithm is the most effective algorithm out of the three for real-time texting over social media applications as it provides comparatively high accuracy, precision, and recall scores in a short time. In the future, our work can be extended with the processing of video and audio messages for the detection of cyberbullying content. The proposed model can be tested on multilingual text in the future. We will also like to

Classifier	Test Size	Accuracy	Conf. Matrix	F1-score	Precision	Recall	Time
SVM	10%	85.857%	TP:968, FN:247 FP:36, TN:750	0.841	75.226%	0.954	154.218 s
SVM	20%	85.854%	TP:1955, FN:474 FP:92, TN:1480	0.839	75.742%	0.941	117.936 s
SVM	30%	83.836%	TP:2893, FN:769 FP:201, TN:21368	0.815	73.547%	0.914	81.920 s
NB	10%	66.317%	TP:546, FN:669 FP:5, TN:781	0.698	53.862%	0.994	5.383 s
NB	20%	64.559%	TP:1052, FN:1377 FP:41, TN:1531	0.683	52.648%	0.974	5.305 s
NB	30%	62.956%	TP:1536, FN:2126 FP:97, TN:2242	0.668	51.328%	0.958	4.700 s
KNN	10%	83.858%	TP:920, FN:295 FP:28, TN:758	0.824	71.985%	0.964	1.089 s
KNN	20%	83.904%	TP:1864, FN:565 FP:79, TN:1493	0.822	72.546%	0.950	1.685 s
KNN	30%	80.886%	TP:2687, FN:975 FP:172, TN:2167	0.791	68.969%	0.926	2.255 s

**Table 1.** Experimental Results on different model configuration

conduct more experimental research to observe if the performance of the KNN classifiers stays the same for a wide range of datasets, or if it varies as per the size, nature, and language of the datasets. We would also like to explore the reason behind such observations. This will further help in exploring the limitations of the KNN classifier as compared to other classification techniques, providing a useful insight into the domain.

**Implementation Code:** The Kaggle link for the implemented code is available at the following link: <https://www.kaggle.com/janab93/spacsec-2021>

## References

1. Sinha, B. B., Dhanalakshmi, R., & Regmi, R. (2020). TimeFly algorithm: a novel behavior-inspired movie recommendation paradigm. *Pattern Analysis and Applications*, 23(4), 1727-1734.
2. Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, 9, 103541-103563.
3. Sinha, B. B., & Dhanalakshmi, R. (2021). Building a fuzzy logic-based McCulloch-Pitts Neuron recommendation model to uplift accuracy. *The Journal of Supercomputing*, 77, 2251-2267.
4. Farley, S., Coyne, I., & D'Cruz, P. (2021). Cyberbullying at work: Understanding the influence of technology. *Concepts, Approaches and Methods*, 233-263.

5. Mladenović, M., Ošmjanski, V., & Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. *ACM Computing Surveys (CSUR)*, 54(1), 1-42.
6. Brochu, M. R. (2017). Cyber bullying: A quantitative study on the perceptions and experiences of female Emirati University students.
7. Noll, H. (2016). *Cyberbullying: Impacting Today's Youth*.
8. Connolly, J., Hussey, P., & Connolly, R. (2014). Technology-enabled bullying and adolescent resistance to report: The need to examine causal factors. *Interactive Technology and Smart Education*.
9. Romsaiyud, W., na Nakornphanom, K., Prasertsilp, P., Nurarak, P., & Konglerd, P. (2017, February). Automated cyberbullying detection using clustering appearance patterns. In *2017 9th International Conference on Knowledge and Smart Technology (KST)* (pp. 242-247). IEEE.
10. Ting, I. H., Liou, W. S., Liberona, D., Wang, S. L., & Bermudez, G. M. T. (2017, October). Towards the detection of cyberbullying based on social network mining techniques. In *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)* (pp. 1-2). IEEE.
11. Rasel, R. I., Sultana, N., Akhter, S., & Meesad, P. (2018, September). Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach. In *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval* (pp. 37-41).
12. Silva, Y. N., Rich, C., & Hall, D. (2016, August). BullyBlocker: Towards the identification of cyberbullying in social networking sites. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1377-1379). IEEE.
13. Isa, S. M., & Ashianti, L. (2017, November). Cyberbullying classification using text mining. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 241-246). IEEE.
14. Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012, June). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 656-666).
15. Sugandhi, R., Pande, A., Chawla, S., Agrawal, A., & Bhagat, H. (2015, December). Methods for detection of cyberbullying: A survey. In *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 173-177). IEEE.
16. Dani, H., Li, J., & Liu, H. (2017, September). Sentiment informed cyberbullying detection in social media. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 52-67). Springer, Cham.
17. Sinha, B. B., & Dhanalakshmi, R. (2019). Evolution of recommender system over the time. *Soft Computing*, 23(23), 12169-12188.
18. Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops (Vol. 2, pp. 241-244)*. IEEE.