Jagannath University Journal of Science
Volume 07, Number II, June 2021, pp. 29–34
https://jnu.ac.bd/journal/portal/archives/science.jsp
ISSN 2224-1698

JnUJSci

Cyberbullying Detection on Social Media Platform: Machine Learning Based Approach

Research Article

Md. Manowarul Islam[1*], Md. Ashraf Uddin[1], Rubaia Rahman[2], Arnisha Akhter[1], Uzzal Kumar Acharjee[1]

[1]*Dept. of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh*
[2]*School of Education, Bangladesh Open University, Gazipur, Bangladesh.*

Received: 20 December 2020
Accepted: 04 March 2021

Abstract : The rapid growth of online communities on social media has gained considerable attractions from researchers and academician. A growing number of human communications largely depends on Internet services, and advances in mobile and network technology have created a path to connect and stay with each other. Unfortunately, amplification of social connectivity also introduces the negative aspects of society leading to bad phenomena like harassment, cyberbullying, and cybercrime. The goal of this study is to design a framework for the detection of bullying posts using natural language processing and machine learning. Different datasets of different social media like Facebook, twitter have been collected and analyzed using machine learning approaches to detect bullying comments or posts such as racism, hate speech, and personal attacks.

Keywords: *Cyberbullying • Cybercrime • Machine learning • Natural language processing • Social media*

1. Introduction

With the development of Internet, the popularity of social media has been increased rapidly over the time and become the most prominent medium of communication for the 21stcentury. However, the rapid growth of social media communication also introduces the negative aspects of society leading to bad phenomena like online harassment, cyberbullying, and cyber-crime (Sameer *et. al.* 2010; Elizabeth *et. al.* 2015). Cyberbullying often leads to severe mental and physical depressions and even suicide attempts especially for women and children. Thus, the detection of bullying text or message has received an increasing amount of attention for the researchers. The detection and prevention of cyber-bullying has not been still extensively explored.

Nowadays, people are using online platform for not only communication but also for business and other means (Ying *et. al.* 2012; Nemanja *et. al.* 2015). People, particularly girls and women have been experiencing online harassment on the social media. Online harassment including bullying, trolling has negative impact on social life. For victims, this kind of behavior can lead to depression and other severe life-threatening problems. This requires a serious attention from the researchers and the cyber-security agencies to control this activity. Technical measurements need to be put in place to monitor and detect potentially harmful online activities. Machine learning (ML) has been extensively explored in social medium and Internet of things (IoT) in

Corresponding Author : Md. Manowarul Islam
E-mail: manowar@cse.jnu.ac.bd

order to detect malware attacks. Most cyberbullying identifying research focuses solely on machine learning techniques which cannot be generalized for detecting abusive words from social media texts because these texts are produced from individuals of diverse language. Therefore, there needs to associate natural language processing with the machine learning techniques.

The current machine learning based works in detecting cyberbullying have one of the three problems. First, exiting works aim only one unique social media (SMP) while training machine learning model. Second, those works focused only one cyberbullying problem at a time. Third, existing works depend on several hand-crafted features of data. Agrawal *et al.* (2018), found that the above bottleneck issues in the identification of cyberbullying can be solved using models based on deep learning. Using three separate datasets, namely Formspring (12k posts), Twitter (16k posts), and Wikipedia (100k posts), the authors performed comprehensive experiments.

The authors (akshi *et. al.* 2019) contributed a review paper that focused on the prospect and implementation of techniques for cyberbullying detection. They analyzed a meta-analytic method to incorporate, explain and critically evaluate the findings of the original studies to clarify new methods to achieve high performance and effective solutions relevant to the established research.

In this paper, we investigate the outcome of using machine learning with natural language processing to identify cyberbullying in social media. Various classification algorithms include Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) classifier have been used to detect the bullying post. Experiment results with two datasets consist of Facebook comments and posts and twits have been explored with the proposal to identify the bulling text.

The rest of the paper is organized as follows: Section 2 presents the relevant literature. In Section 3, the hybrid approach of detecting cyberbullying from social media text is described. In section 4, we discuss the various machine learning algorithms that we use for the experiments. The Section 4 discusses experimental analysis before concluding the paper in Section 5.

2. Related Works

Various studies have been conducted with a conclusion that 10% to 40% Internet users are victims of cyberbullying and the effects of cyberbullying can lead temporary anxiety to suicidal attempts (Sameer *et. al.* 2010; Elizabeth *et. al.* 2015). Over the past few years, several techniques have been proposed to measure and detect offensive or abusive content/behavior on platform like Instagram, YouTube, Yahoo Finance, and Yahoo Answers (Ying *et. al.* 2012; Nemanja *et. al.* 2015). Dinakar *et al.* (2011) proposed a method for the detection of cyberbullying by targeting combinations of profane or negative words. The authors performed affect analysis of small dataset of cyberbullying entries to find out that distinctive features for cyberbullying were vulgar (Michal *et. al.* 2010).

Identification of bullying and harassment in digital networking is a difficult process due to short, noisy and unorganized content where social media users deliberately obfuscate derogatory words or phrases. Inspired by sociological and psychological studies on bullying behavior and its association with sentiments the authors (Harsh *et. al.* 2017) suggested an optimization framework that exploited the sentiment data to accurately detect cyberbullying behavior in social media. They validated their framework using two real world datasets called Twitter and MySpace.

Most research in detecting cyberbullying is mainly supervised learning technique with assumption of the data adequately pre-labelled. However, labelling data is impractical and challenging while data is streaming. The research in (Vinita *et. al.* 2014) recommended a semi-supervised leaning technique to augment samples of training data and utilized a fuzzy SVM algorithm. The training method automatically extracts and extends the training set from the unlabeled streaming text, while learning is carried out as an initial input using a very limited training set.

Although social media offers great opportunities to contact and communicate with all groups of people, it exposes young people to online abuse that is threatening. Recent studies have shown that cyberbullying among young people is a growing problem. Effective security depends on the proper detection of potentially dangerous messages, and smart systems are required to identify potential threats automatically. Van *et al.* (2018) emphasized on automatic identification of cyberbullying from social media texts that include modeling posts, message, and comments written by bullies, victims, and bystanders. For English and Dutch, we identify the compilation and fine-grained annotation of a cyberbullying corpus and carry out a series of binary classification experiments to assess the feasibility of automatic identification of cyberbullying. A linear support vector machines that leverage a rich collection of features and investigate which sources of knowledge contribute most to the mission was used.

The authors (Nijia Lu *et. al.* 2020) focused on identification of textual cyberbullying since the most prevalent type of social media is text. However, incorrect

spellings and symbols, short information, noisy, and unstructured texts affect the efficiency of certain conventional methods. For this purpose, to decide if the text in social media involves cyberbullying, the authors suggested a Char-CNNS (Character-level Convolutionary Neural Network with Shortcuts) model. They used characters as the smallest learning unit, allowing the model to solve spelling errors in real-world businesses and deliberate obfuscation. To identify more bullying signals, shortcuts are used to stitch distinct levels of attributes, and a focal loss mechanism is introduced to solve the issue of class imbalance.

Existing cyberbullying identification researches have concentrated overwhelmingly on the content of conversations, while still ignoring the nature and characteristics of cyberbullying actors. Social research on cyberbullying reveals that the written language used by a harasser varies with the characteristics of the actors, including gender. To train a gender-specific text classifier, the authors (Maral *et. al.* 2012) used a support vector machine model. They demonstrated that considering gender specific language characteristics increases the ability of a classifier to detect cyberbullying to discriminate.

3. Methodology

The proposed framework for detecting cyberbullying depicted in Fig. 1 comprises different modules including collection of raw datasets, Natural Language Processing (NLP), Machine Learning Model and Result Analysis.

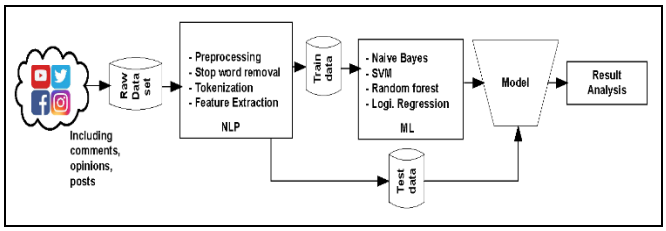


Figure 1. Proposed framework for bully detection

3.1 Data Collection and Text Categorization

In this module, words are corrected with right spelling before passing those to a machine learning classifier to distinguish bullying words from social media posts. In the contexts of culture and countries, some known bullying words might be considered offensive. For, example, "Australia says yes to same-sex marriage" contains the word sex but not a bullying post. Therefore, we need to use an enriched dataset to efficiently identify true bullying attempt from other non-bullying sentences with bullying like words. The data from various online media including Facebook, Twitter has been collected. Figure-1 shows the working principle of the proposal that contains two major modules namely Natural Language Processing (NLP), and Machine Learning (ML).

3.2 Text Preprocessing

This process removes stop-words and whitespace, tokenize words, tag part-of-speech, lemmatizing. This step also includes semantic analysis which analyze the words of the post or comments and outputs a rating based on the number of bullying words. Here, we use correlation based semantic analysis technique. Therefore, natural language processing includes data collection and text categorization, stop word removal, tokenization, and semantic analysis.

3.3 Machine Learning

This module involves in applying various machine learning approaches like Random Forest, Support Vector Machine, Logistic Regression to detect the bullying message and text. The classifier with the highest accuracy is discovered for a particular public cyberbullying dataset. Next section, some common machine learning algorithms is discussed to detect cyberbullying from social media texts.

4. Classifier Model

In this section, we describe the basic of the classifier model that we have used to train and generate the detecting module of the framework.

4.1 Naive Bayes Classifier

The Naive Bayes classifier is a "probabilistic classifier" based on applying Bayes' theorem [14]. Naive Bayes uses conditional probability model, where given a problem instance to be classified, represented by a vector $= (x_1, x_2, x_3 \ldots x_n)$ representing some n features (independent variables), it can calculate the probabilities $P(C|x_1, x_2, x_3 \ldots x_n)$ for each of K possible outcomes or classes C_k. Using Bayes' theorem, the conditional probability can be written as:

$$P(C_k|x) = \frac{P(C_k) \times P(x|C_k)}{P(k)} \quad (1)$$

If each feature xi is conditionally independent of every other feature x_j and for any category C_k, $j \neq i$, then this model can be represented as follows:

$$P(C_k|x_1, x_2, x_3 \ \ldots \ x_n) = P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (2)$$

4.2 Logistic Regression

Logistic regression is a classifier model that uses a logistic function to model a binary dependent variable [15]. Mathematically, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1:

$$P(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad (3)$$

4.3 Support Vector Machines

Support vector machines (SVMs) are powerful supervised machine learning algorithm which can be used for classification [16]. The support vector machine algorithm can find a hyper plane in an n-dimensional space that distinctly classifies the data points where *n* is the number of features. As the text classification problems are linearly separable, Linear SVMs are commonly used for text classification containing a lot of features or attributes. The following equation defines the decision boundary that the SVM returns:

$$f(x) = \quad w^T \; + b \tag{4}$$

Here *w* is the weight vector, *X* is the data dataset to be classified, and *b* is the linear coefficient.

4.4 Random Forest

Random forest is another supervised learning algorithm that can be used both for classification and regression [17]. By constructing a multitude of decision trees, this algorithm can determine the class label of test data very efficiently. The working principle of random classifier can be described by the following stages:

- **Stage 1:** The classifier selects random samples from a given dataset.
- **Stage 2:** The classifier the constructs decision tree from the chosen samples data.
- **Stage 3:** The prediction results of different decision tree are collected.
- **Stage 4:** To select the best results a voting is performed from the prediction results of the decision tree.
- **Stage 5:** Classifier algorithm selects best prediction result from the most voted results.

5. Experimental Results

We have collected Facebook comments from different posts (Dataset-1) and the twitter comments dataset from kaggle.com[18] for (dataset-2). After that, according to our framework, we have applied various machine learning approaches to detect the bullying text and comments. The bullying detection algorithms are implemented using python machine learning packages. What follows we are describing the results of the proposal.

5.1 Performance Metrics

A confusion matrix is a table that represents the number of correct predictions against the number of incorrect predictions shown in table 1.

Table 1. Confusion matrix

	Predicted Class		
Actual Class	Class = Yes	TP (True Positive)	FN (False Negative)
	Class =No	FP (False Positive)	TN (True Negative)

The performances of the proposals are analyzed with respect to the following metrics:

- **Precision:** quantifies the # of positive class predictions that actually belong to the positive class.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

- **Recall:** quantifies the # of positive class predictions made out of all positive examples in the dataset.

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

- **F1-score:** this score takes both false positives and false negatives into account and returns the weighted average of Precision and Recall.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{7}$$

- **Accuracy:** is a measure for how many correct predictions a model made for the complete test dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

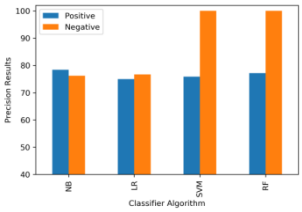
5.2 Results of Dataset-1

At first, we have built this dataset from the user comments on different Facebook posts. Then we classify the data simply two types:

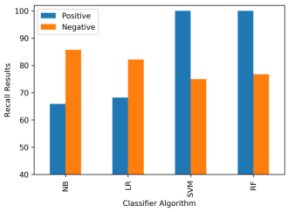
- **Positive:** These types of comments or post are non-bullying comments. For example, the comment like *"A very beautiful movie"* is positive and non-bullying comments.
- **Negative:** This type belongs to bully type comments. For example, *"go away vampire"* is a bullying text or comment and we consider as negative comment.

Figure-2(a),(b), and (c) show the precision, recall and f1-score of the machine learning models. Figure-2(d) shows the overall results on average for all the algorithms while 2(e) shows the performance accuracy. Figure-2(e) shows that both SVM and Random Forest demonstrates higher accuracy than NB and LR classifier where Random Forest results in slightly better accuracy than support

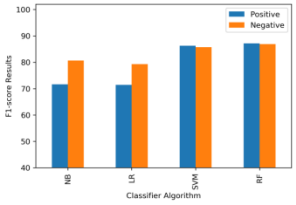
vector machine. Table 2 shows the overall average performance results of the dataset-1.



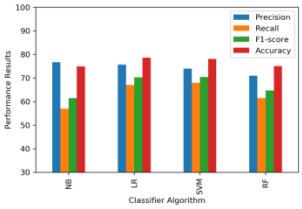
(a) Precision for Dataset-1



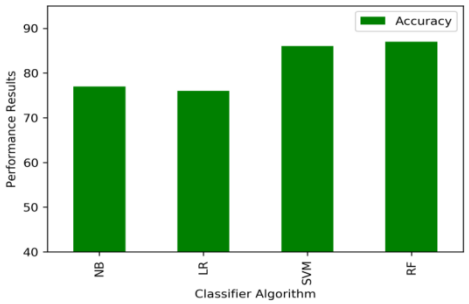
(b) Recall for Dataset-1



(c) F1-score for dataset-1



(d) Average performance



(e) Accuracy for Dataset-1

Figure 2. Performance Results for Dataset-1

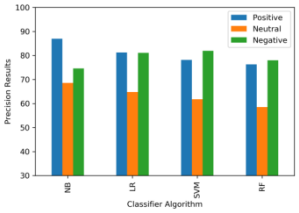
Table 2. Overall Results (Average) for dataset-1.

Method	Precision	Recall	f-1 score	Accuracy
NB	77.28	75.81	76.14	77.00
LR	75.83	75.16	75.37	76.05
SVM	87.93	87.50	85.99	86.00
RF	88.60	88.39	87.00	87.01

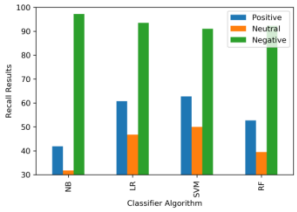
5.3 Results of Dataset-2

In this case, we use the dataset that contains different tweets of the users for different airlines. Here we classify the comments according to 3 categories as follows:

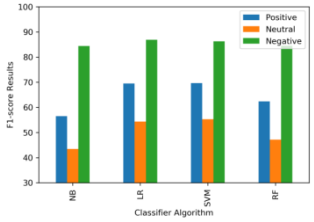
- **Positive:** These types of comments or posts are non-bullying comments. For example, the comment like *"it was amazing, and arrived an hour early."* is positive and non-bullying comment.
- **Neutral**: This type text is simple comments and does not contain anything negative or positive. For example, *"when can I book my flight to Hawaii??"* is just a question not positive or negative expression.
- **Negative:** This type belongs to bully type comments. For example, "you s*ck" is a bullying comment.



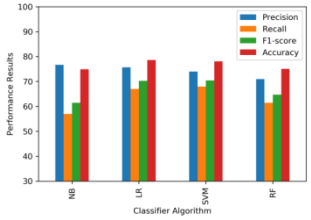
(a) Precision for Dataset-2



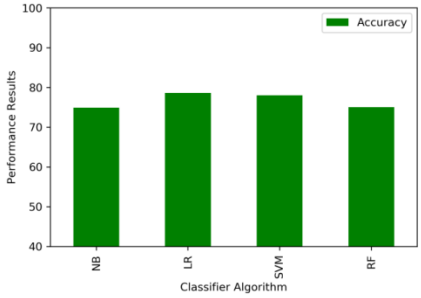
(b) Recall for Dataset-2



(c) F1-score for Dataset-2



(d) Average performance



(e) Accuracy for Dataset-2

Figure 3. Performance Results for Dataset-2

Figures-3(a), (b) and (c) show the precision, recall and f1-score respectively. Figure- 3(d) represents the average performance results of the 4 metrics while Figure-3(e) shows the accuracy of the different classification algorithm. However, although Logistic Regression shows the highest accuracy among all the classifiers for dataset-2, the accuracy difference them is very slight respectively. Table-3 shows the overall average performance summary.

Table 3. Overall Results (Average) for dataset-2

Method	Precision	Recall	f-1 score	Accuracy
NB	76.69	56.97	61.46	74.91
LR	75.69	67.01	70.23	78.62
SVM	73.96	67.95	70.39	78.05
RF	70.94	61.48	64.68	75.05

6. Conclusion

The aim of this study is to investigate the automated identification of posts on social media related to cyberbullying. Manual surveillance for cyberbullying is not feasible considering the overload of information on any sites. Automatic detection of cyberbullying signals can improve moderation and, if necessary, allow rapid response. To some extent, technical methodologies may be able to shield kids from cyberbullying. Yet cyberbullying can be overcome by modifying teenagers' own standards, thinking and actions, and respecting other peers. In combating cyberbullying, teachers and parents could play a significant role. In this paper, we integrated natural language processing with machine learning to get better solutions for cyberbullying identification. Further works are planned to be designed on the detection and classification of cyberbullying text from Bengali texts.

References

S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. Archives of suicide research, vol. 14, no. 3, pp. 206–221, 2010.

E. Whittaker and R. M. Kowalski. Cyberbullying via social media. Journal of school violence, vol. 14, no. 1, pp. 11–29, 2015.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety. In *International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*. pp. 71–80, 2012.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. pp. 29–30,2015.

S. Agrawal and A. Awekar. Deep learning for detecting cyberbullyingacross multiple social media platforms. In *European Conference on Information Retrieval*. pp. 141–153,2018.

A. Kumar and N. Sachdeva. Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimedia Tools and Applications, vol. 78, no. 17, pp. 23 973–24 010, 2019.

K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the Social Mobile Web*. 2011.

M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. In the service of online order: Tackling cyberbullying with machine learning and affect analysis. International Journal of Computational Linguistics Research, vol. 1, no. 3, pp. 135–154,2010.

H. Dani, J. Li, and H. Liu. Sentiment informed cyberbullying detection in social media. *In Joint European conference on machine learning andknowledge discovery in databases*. pp. 52–67,2017.

V. Nahar, S. Al-Maskari, X. Li, and C. Pang. Semi-supervised learning forcyberbullying detection in social networks. In Australasian Database Conference. pp. 160–171, 2014.

C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection of cyberbullying in social media text. PloS one, vol. 13, no. 10, p. 0203794, 2018.

N. Lu, G. Wu, Z. Zhang, Y. Zheng, Y. Ren, and K.-K. R. Choo. Cyberbullying detection in social media text based on character level convolutional neural network with shortcuts. Concurrency and Computation: Practice and Experience, p.5627, 2020.

M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent, 2012.

Naive bayes classifier. https:// en.wikipedia.org/wiki/ Naive Bayesclassifier, accessed: June 2020.

Logistic regression. https://en.wikipedia.org/ wiki/ Logistic regression, accessed: June 2020.

Support vector machine. https://en.wikipedia.org/wiki/ Support vector machine, accessed: June 2020.

Random forest. https://en.wikipedia.org/wiki/Random forest, accessed: June 2020.

Datasets. https://www.kaggle.com/datasets, accessed: June 2020.