I. J. Computer Network and Information Security, 2023, 4, 36-47
Published Online on August 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2023.04.04

A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on FormSpring in Textual Modality

Sahana V.*
JSS Academy of Technical Education Bengaluru/ Department of Information Science and Engineering, Bengaluru, 560060, India
E-mail: sahana26@gmail.com
ORCID iD: https://orcid.org/0000-0002-7768-3642
*Corresponding author

Anil Kumar K. M.
JSS Science and Technology University/ Department of Computer Science and Engineering, Mysuru, 570006, India
E-mail: anilkm@sjce.ac.in
ORCID iD: https://orcid.org/0000-0002-3236-1500

Abdulbasit A. Darem
Northern Border University/ Department of Computer Science, Arar, 9280, Saudi Arabia
E-mail: basit.darem@nbu.edu.sa
ORCID iD: https://orcid.org/0000-0002-5650-1838

Received: 23 February 2022; Revised: 17 June 2022; Accepted: 20 September 2022; Published: 08 August 2023

Abstract: Social media usage has increased tremendously with the rise of the internet and it has evolved into the most powerful networking platform of the twenty-first century. However, a number of undesirable phenomena are associated with increased use of social networking, such as cyberbullying (CB), cybercrime, online abuse and online trolling. Especially for children and women, cyberbullying can have severe psychological and physical effects, even leading to self-harm or suicide. Because of its significant detrimental social impact, the detection of CB text or messages on social media has attracted more research work. To mitigate CB, we have proposed an automated cyberbullying detection model that detects and classifies cyberbullying content as either bullying or non-bullying (binary classification model), creating a more secure social media experience. The proposed model uses Natural Language Processing (NLP) techniques and Machine Learning (ML) approaches to assess cyberbullying contents. Our main goal is to assess different machine learning algorithms for their performance in cyberbullying detection based on a labelled dataset from Formspring [1]. Nine popular machine learning classifiers namely Bootstrap Aggregation or Bagging, Stochastic Gradient Descent (SGD), Random Forest (RF), Decision Tree (DT), Linear Support Vector Classifier (Linear SVC), Logistic Regression (LR), Adaptive Boosting (AdaBoost), Multinomial Naive Bayes (MNB) and K-Nearest Neighbour (KNN) are considered for the work. In addition, we have experimented with a feature extraction method namely CountVectorizer to obtain features that aid for better classification. The results show that the classification accuracy of AdaBoost classifier is 86.52% which is found better than all other machine learning algorithms used in this study. The proposed work demonstrates the effectiveness of machine learning algorithms in automatic cyberbullying detection as against the very intense and time-consuming approaches for the same problem, thereby by facilitating easy incorporation of an effective approach as tools across different platforms enabling people to use social media safely.

Index Terms: Cyberbullying Detection, Machine Learning, Classification, Natural Language Processing, Social Media.

1. Introduction

Social media platforms have grown in popularity as a result of the rapid advancement of Internet technology and currently play a critical part in human life transformation. Social media networks have integrated everyday activities and events like education, entertainment, business, and e-government into human life. Technology usage by Young

people, particularly social media, does, without a doubt, expose them to a variety of psychological and behavioral hazards. One among them is Cyberbullying, a powerful social attack that occurs on social media platforms. Bullying is defined as "an aggressive, intentional act or behavior committed frequently and over time by an individual or a group against a victim who is unable to protect him or herself" [2]. A Cyberbullying is an aggressive act committed using information technologies (IT), such as the Internet [3]. Due to the growing popularity of social media and online communication [4] cyberbullying has been recognized as a major issue [5] and identified as a national health problem [6]. Cyberbullying can be classified into a number of categories as stated in [7,8]:

- Flooding: Bullies comment regularly on the same person or use nonsense comments or even select the enter key to prevent the victim from speaking.
- Masquerade: Using another person's account to bully a victim in a chat room, forum to tarnish the victim's reputation.
- Flaming (bashing): Personal attack is carried out by two or more users. All of the posts contain bullying language and involve a heated, short-lived argument.
- Trolling (baiting): Responds to an emotionally charged thread with comments intended to provoke a war, even though the comments don't accurately reflect the poster's opinion.
- Harassment: Typically involves an assumed victim-bully relationship, which is most reminiscent of conventional bullying. A constant stream of abusive messages is sent to the victim for an extended period of time.
- Cyberstalking and cyberthreats: These include threatening, aggressive, or extortionary messages.
- Denigration: This category involves spreading obscene, negative, or false rumors about other people via forums, chat rooms, or websites.
- Outing: This involves posting private, embarrassing, or confidential information in a public forum or chat room. Unlike denigration, this type of bullying involves a close relationship (online or in person) between the bully and the victim.
- Exclusion: Teenagers and young people have become more likely to engage in this type of cyberbullying by ignoring the victim in chat rooms and conversations.

Furthermore, cyberbullying is associated with negative mental health outcomes such as suicidal ideation, attempted to commit suicide, and emotional and social concerns, as well as despair, anxiety and other forms of self-harm. As a global phenomenon, cyberbullying occurs worldwide, particularly in India, Brazil, and the United States. Indian children have been the most victims of cyberbullying so far in 2018. A study by the First Site Guide web portal [9] conveys that more than 37% of Indian parents claim their children had been victims of cyberbullying at least once as shown in Fig.1. This is up 5% from 2016. Monitoring online content effectively is essential to effective CB detection, but it is practically impossible for moderators to do so manually due to the large volume of information on the Web. To tackle this problem, automatic cyberbullying detection on social media is critical and should be prioritized in order to protect children and society from its negative consequences.

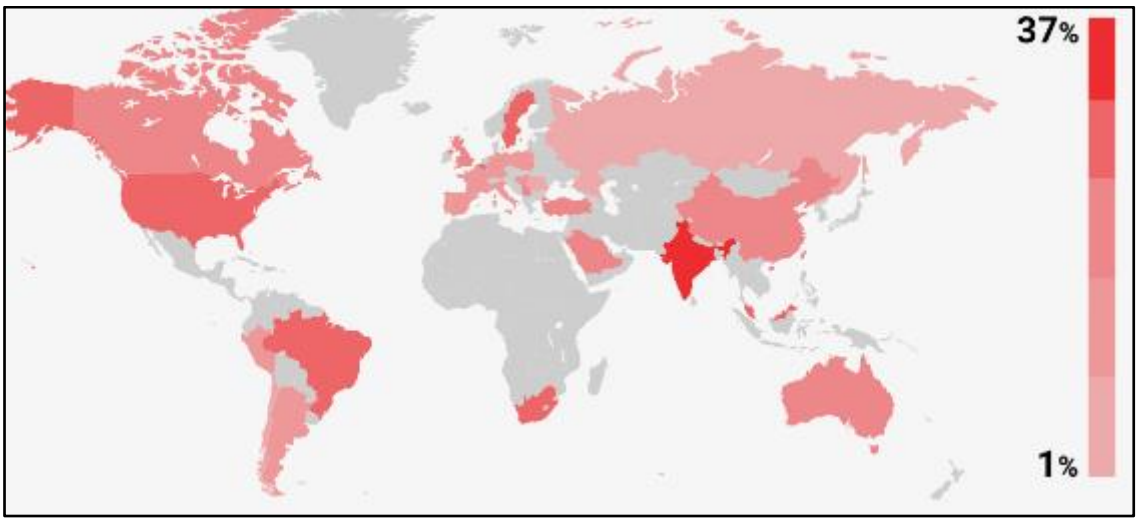

Fig.1. Cyberbullying facts and statistics [9]

As a result, various global initiatives to tackle cyberbullying have been proposed. Kiva [10] for instance, was launched at Turku University in Finland which is an anti-cyberbullying program, France started an anti-harassment campaign [11], and the Belgian government launched an anti-cyberbully project [12]. Machine learning for detecting and classifying offensive language is one of the avenues of cyberbullying research. The research objective of the proposed paper is to explore the potential of various machine learning algorithms used in binary classification of CB

related documents as bullying and non-bullying using effective feature engineering and NLP techniques.

In this context, several machine learning methods have been researched, including Bagging classifier, Random forest classifier, SGD classifier, Linear SVC, Logistic Regression, AdaBoost classifier, Multinomial Naive Bayes (MNB), Decision tree classifier, and K Neighbors classifier. We have conducted experiments on a labelled dataset from FormSpring (Social network based on questions and answers), which was collected and labelled by the authors of [1]. The work's key contributions are as follows:

- Thorough analysis of high-quality articles to discover the most extensively used machine learning (ML) algorithms for detecting cyberbullying on social media platforms (SMPs).
- Experimentation using a real world, Formspring dataset and exploring potential of various ML classifiers.
- Usage of effective feature extraction technique to boost the performance of the classifiers in our automated cyberbullying detection model.
- Comparison of results of nine machine learning classifiers, that are used in detecting cyberbullying. The study reveals the benefits and limitations of machine learning classifiers in text classification related to cyber bullying domain.

The paper is organized in the following manner: A brief overview of relevant works is provided in Section 2. The details of the dataset, NLP techniques, feature extraction, ML algorithms implemented are presented in Section 3. The outcomes of the experiments along with the discussions are presented in Section 4. Conclusion is discussed in Section 5.

2. Related Works

Many machine learning algorithms including both supervised and unsupervised learning algorithms have been used for detection of cyberbullying in social media post. Raisi, E., & Huang, B. (2017) devised a method for detecting cyberbullying that is weakly supervised [13]. They showed that participant vocabulary consistency can lead to the discovery of bullying incidents and novel bullying language. To recognize the sentiment and context of a sentence, a supervised machine learning technique based on a bag-of-words approach was presented by D. Yin et al. [14]. Additionally, they demonstrated that adding sentiment and contextual feature attributes to TFIDF can help detect online harassment. The accuracy of 61.9 % was achieved in this method. Several studies have examined supervised learning techniques focused on detecting bully content on social media. According to Amanpreet Singh and Maninder Kaur (2019), Naive Bayes (NB) and Support Vector Machine (SVM) are the most widely used techniques for achieving effective results in detecting bully content [15]. While most studies deal with bullying in text form, photos and videos can also be used as methods of internet harassment with much greater potential negative consequences. Researchers are assisted in discovering the significant characteristics of content-based Cybercrime detection methods by the methodical analysis work performed by these authors. Mengfan Yao et al. (2019) proposed a method for detecting cyberbullying on Instagram media sessions that is both timely and accurate [16]. Using a sequential hypothesis testing formulation, they proposed a method for classifying comments with a minimal number of features and high levels of accuracy. When compared to other cutting-edge technologies, this cyberbullying detection system attained an accuracy of 0.80.

Homa Hosseinmardi et al. (2016) suggested a predictor that can foresee cyberbullying episodes occurring before they occur [17]. The predictive ability of various features was investigated using a logistic regression classifier. They demonstrated that non-text features including user meta data and image were critical in predicting cyberbullying, with a LR classifier achieving 0.78 precision and 0.72 recall. The rate of false positives was as low as 0.01. The authors of [18] researched on whether studying social network features may help detect cyberbullying more accurately. According to the authors, who examined the social network structure between users and inferred factors like network embeddedness, number of friends and relationship centrality, combining textual information with social network properties might significantly increase the detection of cyberbullying. Sweta Agrawal and Amit Awekar (2018) trained Deep Neural Networks using datasets from FormSpring, Wikipedia, and Twitter, with a particular focus on swear words and their use as task features [19]. They investigated into how the vocabulary for such models differs between different Social Media Platforms. In [1], with the Weka tool package, the labeled data was used to train a computer to identify bullying content using machine learning techniques. With an instance-based learner and C4.5 decision tree learner, they were able to detect 78.5 % of true positives.

Dinakar et al. (2012) proposed an approach for detecting and reducing cyberbullying [20]. Their research had a broader scope, as it featured not only ways for detecting cyberbullying, but also solutions for resolving the issue. In comparison to earlier study, this was an improvement. In an English dataset, their classifiers scored between 58 and 77 % F-score. The outcomes differed based on the type of harassment they were trying to categorize. SVM was again the best classifier they proposed, confirming the efficiency of SVMs in cyberbullying detection, comparable to study carried out in 2010 using a Japanese dataset [21]. Psychological features such as personalities, sentiments, and emotions were incorporated by Balakrishnan, Khan, and Arabnia [22] and Rosa et al. [23] to improve automatic cyberbullying detection. The usage of personalities and sentiments increased cyberbullying detection, while emotion had no such effect. Extraversion, neuroticism, agreeableness, and psychopathy had greater impact in detecting cyberbullying than other personality qualities, according to a further examination of the personalities. A mobile application developed by

the authors of [24] employs a machine learning model to detect cyberbullying in children among parents. Using the labelled dataset on account creation dates according to time period, Chatzakou et al. (2017) analyzed the Twitter user accounts in the dataset for two time periods [25]. According to their study, 38% of users who were identified as bullies earlier deleted their Twitter accounts later. As the reason for the deletion, they claimed that they wanted to prevent Twitter from suspending their accounts for spam, fake, and abusive behavior. Bullies might also benefit from the deletion because they can hide their identities. It has been reported that bullies tend to create their accounts later than normal users, according to Ribeiro et al. (2017). Rather than using their real accounts to cyberbully, bullies create other accounts to camouflage their real identities [26]. A later account will be deleted after a certain period of time. Therefore, a person's account creation date can provide useful information about their online bullying behavior.

We have conducted a survey on research efforts related to CB detection. According to this survey, CB prevention has received growing attention in recent years. The majority of studies are based on supervised learning methods, but researchers have shown willingness to incorporate emerging work from other fields of NLP in order to improve performance. Accuracy, precision, recall, and F1 score are key metrics used in evaluating classifiers. Several papers within our sample described these metrics as part of their experiments, but the comparison of those studies is not straightforward based on those metrics. The datasets used by the studies will directly influence the results of the study. It is meaningless to compare the achieved metrics' values without conducting the experiments on the same dataset. So, in the proposed work, along with using nine machine learning algorithms and effective feature extraction technique for cyberbullying detection, we have also compared our results with the work of few researchers, who have used the same dataset for their experiments.

3. Methodology

This part discusses about the FormSpring dataset used for CB detection, its visualization, and the proposed model for cyberbullying detection including details about pre-processing, feature extraction, and classification algorithms.

3.1. Dataset

Theoretical and practical challenges exist in cyberbullying detection in social media utilizing keywords associated with cyberbullying and applying MLs for detection. From a practical standpoint, the researchers are still using the learning model to detect and classify offensive content. However, constructing an efficient and effective cyberbullying detection algorithm still faces significant challenges in terms of classification accuracy and model implementation. Kelly Reynolds et al. [1] collected and labeled cyberbullying data from Formspring. We chose the Formspring dataset since it was mostly populated by teens and college students, and its data contains a high proportion of bullying content due to the anonymity of the entries. A number of 12857 conversations are contained in this dataset that are questions and their answers which are annotated as either cyberbullying or not. There are two segments in the data set. 70% of the posts used for training are in the first segment, whereas 30% of the posts used for prediction are in the second segment. The texts or responses were divided into two categories:

- Non-bullying Text: These are pro or positive remarks or posts. For example, positive and non-bullying comments like "This photo is extremely beautiful".
- Bullying Text: This category refers to bullying or harassment comments. "Go away bitch," for example, is a bullying text or message that we regard to be a negative comment.

3.2. Cyberbullying Detection Model

Fig. 2 depicts the proposed model for cyberbullying detection, which is divided into four phases: preprocessing, feature extraction, classification, and evaluation.

A. Pre-processing

In order to detect cyberbullying, the preprocessing stage is critical. Various superfluous characters or text can be found in real-world posts or messages. Numbers and punctuation, for example, have no impact on the detection of bullying. The comments must be cleaned and prepared for the detection stage before applying the machine learning techniques to them [27]. Various processing tasks are performed in this step, including the elimination of any characters that are not relevant such as punctuation, stop-words, numerals, tokenization, stemming, and so on. It includes both text cleaning and the removal of spam content. In the proposed model, it was utilized to decrease and remove undesired noise in text detection. As a result of this preprocessing, the remaining words have been stemmed back to their original roots and a cleaned dataset has been prepared to be run and evaluated by the proposed model.

B. Feature Extraction

In the text classification process, feature extraction is an important stage and so in cyberbullying detection as well. To reduce dimensionality, feature extraction is employed in machine learning after preprocessing. The Count Vectorizer method was utilized to extract features in the proposed methodology. CountVectorizer is an excellent Python tool from the scikit-learn toolkit. Here, the text is converted into a vector based on the frequency /count of each word in the text. It

is beneficial to convert each word into a vector when working with a large number of such texts (to be used in future text analysis). A vectorized representation of the features is created (in a well-arranged manner).

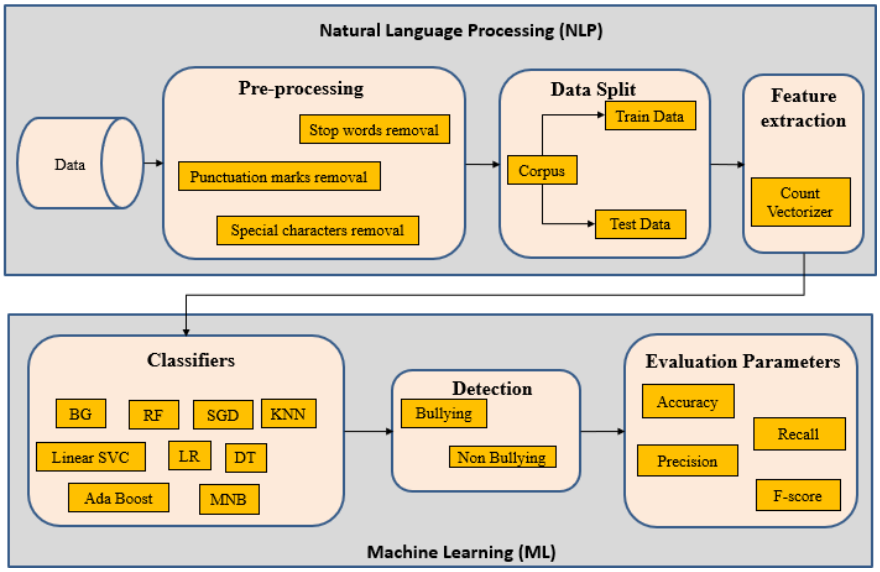

Fig.2. Overview of cyberbullying classification pipeline

C. Machine Learning

Another method of detecting cyberbullying is to utilize machine learning-based cyberbullying keywords, which has been widely utilized by various studies. Furthermore, artificial intelligence based on machine learning allows a computer to learn and improve without being programmed [28]. ML is divided into three categories: supervised, semi-supervised, and unsupervised algorithms. To develop a model that predicts the expected outcome (i.e., based on data with annotations or labels), several training examples in supervised algorithms are used. Unsupervised algorithms, on the other hand, are not data-driven and are primarily used for clustering [29].

Given CB is a classification problem (i.e., classifying an occurrence as bullying or non- bullying), this study used different supervised learning algorithms to enhance the accuracy of their classification and performance in identifying CB on social media. The following are the classifiers used in this study:

- *Bagging classifier*

Bootstrap aggregating (Bagging) is an old and powerful method in ensemble learning that builds various classifiers by learning an ensemble of classifiers over bootstraps. It is a meta-estimator classifier in which different base classifiers are fitted to random subsets of original dataset and the averaged or voted predictions are combined to produce a final prediction. It is constructed by introducing randomization into its construction. The base classifiers are trained in parallel, each with a training dataset derived from original training dataset by randomly replacing N data points with new ones. For each base classifier, the training set is independent. Through voting or averaging, bagging reduces overfitting [30].

This algorithm includes classifier generation and the classification:
Classifier generation:
a. Let the training set size be N.
b. For each iteration t:
i. Sample N occurrences with replacement from the initial training set
ii. Learning algorithm is applied to the sample
iii. Save the final classifier
Classification:
a. For each classifiers t:
i. Utilize a classifier to predict the instance's class
b. Return the most predicted class. [31]

- *Random Forest*

This classifier can be constructed by combining multiple decision trees [32]. Trees provide class predictions independently. The ultimate output is the maximum number of classes predicted. It is created by combining numerous

decision trees into one using an accurate learning model based on supervised learning. Based on the majority of votes, the RF determines the final outcome using the predictions from each tree, as it is preferable to use multiple decision trees rather than one. For instance, if there are two classes, say A and B, and the majority of decision trees predict the class label B of some instance, Random Forest will choose the class label B as in (1)

$$f(x) = B \text{ is the majority vote of all trees} \tag{1}$$

- *Stochastic Gradient Descent optimizer (SGD)*

Because of its simplicity, fast convergence, and applicability for non-convex functions, SGD is regarded the default standard optimization technique for many gradient-based ML classification models such as neural networks and logistic regression. In [14, 33, 34], SGD was utilized to create cyberbullying prediction models on social networking platforms. SGD updates the parameters on each example (x^i, y^i) rather than going through all the samples. Therefore, learning happens on every example, as in (2)

$$\theta = \theta - \eta . \nabla_\theta J(\theta; x^i; y^i) \tag{2}$$

- *Linear SVC*

The Linear SVC technique uses a linear kernel function to classify data and finds that it is very effective for large datasets. When compared to SVC models, Linear SVC models have additional parameters, such as penalty normalization (L1 or L2) and loss function. Linear SVC is dependent on the linear kernel methodology and, as a result, cannot be changed. The following are advantages of linear SVC:

i. It allows for the use of several regularisation techniques in the formulation.
ii. SVM uses relatively little memory.
iii. On the test error rate of SVC, there is an approximate bound.
iv. Whenever there is a definite margin of distinction between classes, it performs considerably better.
v. Linear classifier is faster compared to non-linear classifier.

- *Logistic Regression*

LR is a statistical model which determines class probabilities using a sigmoid logistic function as opposed to a straight line. In order to fit logistic regression models, maximum likelihood estimation is frequently performed. It determines which class an input belongs to by providing a probability value between 0 and 1. As a probability function, the sigmoid function [35] is used to model the output of a problem as in (3)

$$sig(x) = \frac{1}{\{1+\exp(-x)\}} \tag{3}$$

$$A = LT + C \tag{4}$$

$$T(X) = sig(A) \tag{5}$$

The classifier's hypothesis function is assumed by T(x), weights are calculated by L, bias is calculated by C, and T is vector of features (input) as in (4) and (5). It is assumed that class is 1 if h(x)>0.5; otherwise, it is assumed that class is 0. Due to below benefits, we choose to employ LR:

i. Implementation and interpretation of LR are simple, and training is quick and efficient.
ii. It classifies unknown records very quickly. Although it is less likely to overfit, logistic regression can overfit in high-dimensional datasets.

- *AdaBoost classifier*

AdaBoost [36] and Extreme Gradient Boosting (XGBoost) [37], which is a more extended and optimized variant, are examples of boosting algorithms. Weak learners, such as decision trees, augment previous residuals over time as they are incrementally built. Averaging the outcomes of random predictors is more effective compared to generating random predictors (RF), so, members cast a weighted vote. AdaBoost classifier implementation is illustrated in Fig. 3, where a similar kind of dataset with two classes and two features is shown where weak learner #2 improves the accuracy of the misclassified observations by making a mistake committed by weak learner #1. When the two-weak classifiers are combined (strong leaner), observations misclassified are further refined to improve accuracy.

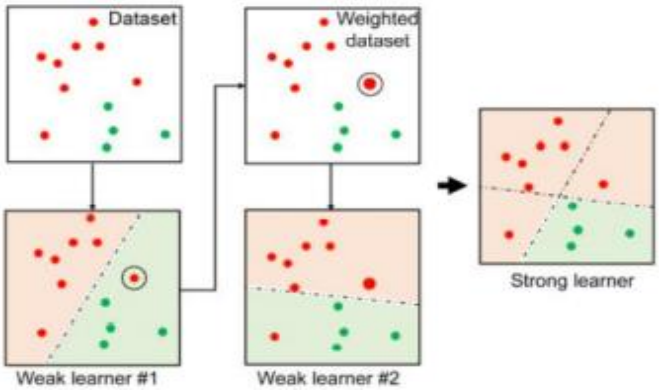

Fig.3. Implementation of adaboost classifier

- *Decision Tree*

Both classification and regression can be done with the DT classifier [38,39]. It can be used to represent and also to make a decision. Each core node represents a condition, and each leaf node represents a conclusion in a decision tree. A classification tree shows which class the target belongs to. A regression tree is used to predict the expected value for an addressed input.

- *Multinomial Naive Bayes (MNB)*

In text categorization, Multinomial Naive Bayes (MNB) has been frequently employed. MNB frequently employs a parameter learning method known as Frequency Estimate (FE), which estimates word probabilities by estimating suitable frequencies from data, given a set of labeled data [40]. By using the Bayes theorem among features, NB classifiers were created. Using training data, Bayes-optimal parameter estimates are determined for a parametric model that produces the text. Using these approximations, the data is classified. In addition to continuous functions, NB classifiers can handle categorical functions as well. With distinct functions, it is possible to estimate one-dimensional kernel density in place of high-dimensional density. With strong (naive) assumptions of independence, the NB algorithm uses the Bayes theorem to learn. We considered MNB because of the following significant rare features:

i. The user simply needs to calculate probability, therefore it is very simple to implement.
ii. MNB can be applied on both continuous and discrete data.
iii. It is relatively easy to use and can be applied to predict real-time applications.
iv. Large datasets may be handled with ease and its scalability is excellent.

- *k-Nearest Neighbors classifier (kNN)*

As input, kNN uses a majority vote to categorize the input samples based on the training data with k-closest training samples. It is frequently used in conjunction with Naive Bayes as a baseline. The classifier is quick and easy to train, but it is extremely vulnerable to outliers and overfitting. It is k in the kNN, that determines if the new sample belongs to the same class as its nearest neighbors. In this study, neighboring samples are categorized into two main categories, namely cyberbullying and non-cyberbullying. Our study used k = 1 configuration, the simplest form of the kNN algorithm, in which input samples are simply assigned to the class of the first nearest neighbor [41]. We determined the best k by comparing the mean errors in the model's predictions with the labeled test data, using a method known as the elbow method.

4. Experiments and Results

Nine machine learning classifying algorithms were applied to the training dataset to categorise it, in order to compare and contrast the effectiveness of those algorithms. Once the classifiers have been trained, the testing dataset was pre-processed, its features were extracted, and was run through the classifiers to detect the polarity. Then, the polarity was used to compare the classifiers' accuracy. We conducted binary classification studies for the automatic detection of cyberbullying using a number of machine learning techniques by making use of Scikit-learn. The open source Python library NumPy (v1.21.5), is used to work with arrays, has been imported as a part of our experimental settings. Additionally, we have utilized Pandas (v0.23.2), an open-source Python-based framework for data analysis and manipulation. Pandas is an ideal tool for processing this complicated real-world data and it is built on NumPy. It also supports importing data from a different file types, including comma-separated values, JSON, SQL, and Microsoft Excel. In addition to cleaning and wrangling data, Pandas supports operations such as merging and restructuring.

Additionally, we used Matplotlib, a Python tool for cross-platform data visualisation and graphical charting. The classifiers were then imported via the Sklearn library and assessed using the following evaluation metrics.

4.1. Evaluation Metrics

In this study, the usefulness of the model was evaluated using popular evaluation metrics and to see how well the suggested model distinguishes cyberbullying from non-cyberbullying text documents. As mentioned earlier, nine machine learning algorithms have been used in this study. A brief description of metrics are as follows:

A. Accuracy

One of the simplest performance metrics is accuracy, which simply represents the proportion of correctly predicted occurrences to all occurrences as in (6).

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \tag{6}$$

Where,

- TP (True Positive) is the percentage of cyberbullying instances correctly classified as cyberbullying.
- TN (True Negative) is the percentage of non-cyberbullying instances correctly classified as non-cyberbullying.
- False positive (FP) is the percentage of non-cyberbullying instances incorrectly classified as cyberbullying.
- False negative (FN) is the percentage of cyberbullying instances incorrectly classified as non-cyberbullying.

B. Precision

Precision is calculated by dividing the number of true positives (TP) by the number of true positives plus false positives (FP) as in (7).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \tag{7}$$

C. Recall

Calculating recall involves dividing the number of true positives (TP) by the number of true positives plus false negatives (FN) as in (8).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{8}$$

D. F-Measure

The F1-score, which is defined as the harmonic mean of precision and recall, is also related to these variables as in (9).

$$\text{F1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{9}$$

4.2. Results

The findings of the experiments are presented along with a discussion of their significance. First, each classifier's performance results have been recorded in Table 1. It also shows each classifier's evaluations in terms of accuracy, precision, recall, and F1 score, during both training and testing time. Second, Table 2 shows the training and testing time complexity of each algorithm used in the experimentation.

Table 1. Performance summary of machine learning algorithms

Algorithm	Accuracy: Test	Precision: Test	Recall: Test	F1 Score: Test	Accuracy: Train	Precision: Train	Recall: Train	F1 Score: Train
Bagging Classifier	0.854	0.603	0.383	0.468	0.980	0.991	0.890	0.938
Random Forest Classifier	0.859	0.642	0.364	0.465	1.000	1.000	0.999	0.999
SGD Classifier	0.843	0.549	0.392	0.457	0.977	0.983	0.880	0.928
Linear SVC	0.835	0.512	0.397	0.447	0.990	0.993	0.949	0.970
Logistic Regression	0.862	0.694	0.327	0.445	0.939	0.975	0.646	0.777
AdaBoost Classifier	0.865	0.729	0.318	0.443	0.870	0.747	0.327	0.455
Decision Tree Classifier	0.818	0.455	0.416	0.434	1.000	1.000	0.999	0.999
Multinomial NB	0.852	0.628	0.294	0.401	0.900	0.789	0.538	0.640
Kneighbors Classifier	0.838	0.581	0.126	0.207	0.852	0.726	0.171	0.277

Table 2. Time complexity of machine learning algorithms

Algorithm	Prediction Time	Training Time
Bagging Classifier	0.322	37.237
Random Forest Classifier	3.496	17.820
SGD Classifier	0.002	0.087
Linear SVC	0.001	0.597
Logistic Regression	0.000	0.485
AdaBoost Classifier	0.187	0.864
Decision Tree Classifier	0.045	5.018
Multinomial NB	0.016	0.016
KNeighbors Classifier	6.494	0.016

We found that the AdaBoost classifier has got the best accuracy and precision of 0.865 and 0.729 respectively when compared to other machine learning algorithms. Using AdaBoost classifier, the accuracy is 0.865, meaning the model correctly predicted 86% of the comments labeled as cyberbullying or non-cyberbullying. This classifier displays a precision score of 0.729, indicating that 72% of comments that it predicted as cyberbullying/non-cyberbullying belong to their respective categories. This classifier also exhibits a recall score of 0.318. This simply implies that the classifier was able to find 31% of the comments the classifier predicted as cyberbullying/non-cyberbullying in the pool of comments. As a weighted average of precision and recall for cyberbullying / non-cyberbullying text, the model showed an F1-score of 0.443 (44%). It is observed that even the best performing AdaBoost classifier has a low recall (<0.5), meaning, there are a high number of false negatives produced by the classifier as a result of imbalanced classes. For this reason, we must prepare our data before handling an imbalanced class problem by over-sampling or under-sampling.

Fig. 4 shows the performance summary of machine learning algorithms used in our proposed work. We found that the training time for the Multinomial Naive Bayes and k-Nearest Neighbors classifier being lowest 0.015 and that for Bagging classifier being highest 37.23.

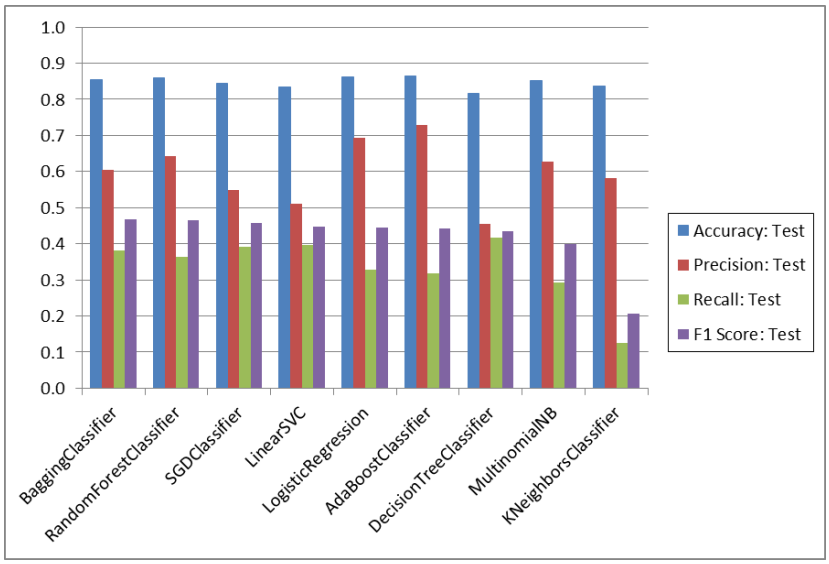

Fig.4. Performance summary of machine learning algorithms

In addition to the previous experiments, we evaluated and compared our best performing classifiers with the proposed work of Kelly Reynolds et al. [1], Sweta Agarwal et al. [19], Vikas S Chavan et al. [42], who have used the same dataset. The summary of results is shown in Table 3. In [1], authors used C4.5 decision tree learner and an instance-based learner and were able to identify the true positives with 78.5% accuracy. In our model, the decision tree has obtained 81.8% accuracy. Sweta Agarwal et al. [19] have used Random Forest classifier to detect cyberbullying content and have obtained an F1-score as 0.298. In our model, using RF, we have obtained 0.465 as F1-score. Vikas S Chavan et al. [42] have used LR and have obtained precision of 0.64, whereas in our model, using LR, we have obtained precision of 0.69. Fig. 5 shows comparison of current results with related previous work. Hence, it is found that our proposed model outperforms all other classifiers and is ranked as the best results in terms of Accuracy, F1-Score and Precision.

Table 3. Comparison with related previous work

	Classifier	Precision	F1 score	Accuracy
Kelly Reynolds (2011)	DT	-	-	0.785
Sweta Agarwal (2018)	RF	-	0.29	-
	NB	-	0.36	-
Vikas S Chavan (2015)	LR	0.64	-	0.737
Current Results	LR	**0.69**	0.45	**0.862**
	DT	0.45	0.44	**0.818**
	MNB	0.63	**0.40**	0.851
	RF	0.64	**0.46**	0.858

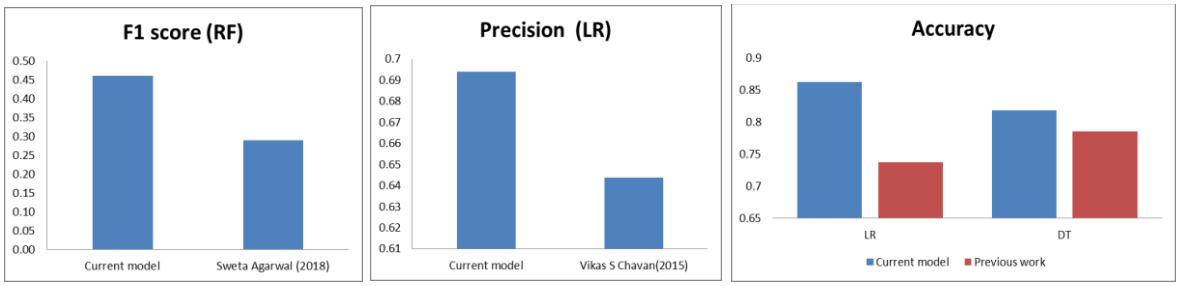

Fig.5. Comparison of current results with related previous work

5. Conclusions

With the increasing prominence of social media platforms and growing usage of social media by youths, cyberbullying is becoming increasingly common, and it is causing significant social problems. To avoid the negative impacts of cyberbullying, an automatic cyberbullying detection system must be developed. Given the importance of cyberbullying detection, we explored the automatic detection of postings on social media platforms related to cyberbullying using the Count Vectorizer function in this study. To identify bullying text, nine machine learning methods are utilized, and we discovered that the AdaBoost classifier surpasses the others with the best accuracy 86.5%. Also, Bagging classifier outperformed the other classifiers with the best F1 score 0.468. In addition, we compared our results with three related works that used the same dataset and discovered that our model surpassed their classifiers in terms of accuracy, f-score, and precision. Our approach will improve cyberbullying detection by attaining high accuracy in contrast to previous methods, allowing people to use social media safely. The results of the current evaluation will aid future researchers in selecting a classifier that is appropriate and sufficient for the cyberbullying datasets because modifications are required to further boost classification accuracy. Our ultimate goal in detecting cyberbullying on social media is to flag as many online risks as we can, hence minimising the need for manual social media patrolling. We hope to achieve this by concentrating on further improving our recall and precision. In the future, we'll strive to collect a variety of datasets to assess our model's performance, as detecting cyberbullying pattern is limited by the size of training data. Proposed cyberbullying detection model relies on binary classification (bullying or non-bullying), thus our future research may take a multi-class classification method.

References

[1] 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. DOI: 10.1109/ICMLA.2011.152

[2] Olweus, D. (1994). Bullying at school. In *Aggressive behavior* (pp. 97-130). Springer, Boston, MA. DOI: 10.1007/978-1-4757-9116-7_5

[3] Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, *49*(2), 147-154. DOI: 10.1111/j.1467-9450.2007. 00611.x

[4] Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., ... & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access*, *7*, 70701-70718.

[5] O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Pediatrics*, *127*(4), 800-804. DOI: 10.1542/peds.2011-0054

[6] 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).

[7] S. Nadali, M.A.A. Murad, N.M. Sharef, A. Mustapha, S. Shojaee, "A review of cyberbullying detection: An overview," International Conference on Intelligent Systems Design and Applications, ISDA, 325–330, 2014, DOI:10.1109/ISDA.2013.6920758

[8] Willard, "Parent Guide to Cyberbullying and Cyberthreats," 1–14, 2014.

[9] Ogi Djuraskovic, 2022, Cyberbullying Statistics, Facts, and Trends with Charts. https://firstsiteguide.com/cyberbullying-stats/

[10] Mc Guckin, C., & Corcoran, L. (Eds.). (2017). *Cyberbullying: where are We Now?: A Cross-national Understanding*. MDPI.

[11] Vaillancourt, T., Faris, R., & Mishna, F. (2017). Cyberbullying in children and youth: Implications for health and clinical practice. *The Canadian journal of psychiatry*, *62*(6), 368-373. DOI: 10.1177/0706743716684791

[12] Görzig, A., & Ólafsson, K. (2013). What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries. *Journal of Children and Media*, *7*(1), 9-27.

[13] Raisi, E., & Huang, B. (2017, July). Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 409-416).

[14] 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.

[15] Singh, A., & Kaur, M. (2019). Content-based cybercrime detection: A concise review. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(8), 1193-1207.

[16] Yao, M., Chelmis, C., & Zois, D. S. (2019, May). Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference* (pp. 3427-3433). DOI: 10.1145/3308558.3313462

[17] Hosseinmardi, H., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2016, August). Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 186-192). IEEE. DOI: 10.1109/ASONAM.2016.7752233

[18] Huang, Q., Singh, V. K., & Atrey, P. K. (2014, November). Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-aware Multimedia* (pp. 3-6). DOI: 10.1145/2661126.2661133

[19] Agrawal, S., & Awekar, A. (2018, March). Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval* (pp. 141-153). Springer, Cham. DOI: 10.1109/ICMLA.2011.152

[20] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), 1-30. DOI: 10.1145/2362394.2362400

[21] Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., & Araki, K. (2010). Machine learning and affect analysis against cyber-bullying. *the 36th AISB*, 7-16.

[22] Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, *90*, 101710.DOI: 10.1016/j.cose.2019.101710

[23] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*, 333-345. DOI: 10.1016/j.chb.2018.12.021

[24] Thun, L. J., Teh, P. L., & Cheng, C. B. (2022). CyberAid: Are your children safe from cyberbullying? Journal of King Saud University-Computer and Information Sciences, 34(7), 4099-4108. DOI: 10.1016/j.jksuci.2021.03.001

[25] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference, 13–22. DOI: 10.1145/3091478.3091487

[26] Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W. (2017). "Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. Available at: http://arxiv.org/abs/1801.00317.

[27] Fosler-Lussier, E., Riloff, E., & Bangalore, S. (2012, June). Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[28] Dinakar, K., Reichart, R., & Lieberman, H. (2011, July). Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*.

[29] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012, September). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 71-80). IEEE. DOI: 10.1109/SocialCom-PASSAT.2012.55

[30] Aziz S., M. U. Khan, Z. Ahmad Choudhry, A. Aymin and A. Usman, "ECG-based Biometric Authentication using Empirical Mode Decomposition and Support Vector Machines," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2019, pp. 0906–0912. DOI: 10.1109/IEMCON.2019.8936174.

[31] Dey, D. (2018). ML | Bagging Classifier.

[32] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*(1), 217-222. DOI: 10.1080/01431160412331269698

[33] Bayzick, J., Kontostathis, A., & Edwards, L. (2011). Detecting the presence of cyberbullying using computer software.

[34] Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433-443. DOI: 10.1016/j.chb.2016.05.051

[35] Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., and Castelnuovo, B. (2018). A comparative study of logistic regression-based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. BMC medical informatics and decision making, 18(1), 77. DOI: 10.1186/s12911-018-0659-x.

[36] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

[37] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). DOI: 10.1145/2939672.2939785

[38] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer. DOI: 10.1007/978-1-4614-7138-7

[39] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674. DOI: 10.1109/21.97458
[40] Su, J., Shirab, J. S., & Matwin, S. (2011, January). Large scale text classification using semisupervised multinomial naive bayes. In *ICML*.
[41] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-39964-3-62
[42] Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2354-2358). IEEE. DOI: 10.1109/ICACCI.2015.7275970

Authors' Profiles

Sahana V. is working as Assistant Professor in the Department of Information Science and Engineering at JSS Academy of Technical Education, Bengaluru. She has completed her M.Tech. in 2015 from Department of Computer science and Engineering, RNSIT, Bengaluru under the Visveswaraya Technological University. She is pursuing her Ph.D. in Web mining at JSS Science and Technology University, Mysuru.

Dr. Anil Kumar K. M. is working as Professor and Associate Dean (Ranking, Accreditation and Analytic) in Computer Science and Engineering department of JSS Science and Technology University, Mysuru. He has total experience of 24 years. He has completed his post-doctoral from Deakin University, Australia. His areas of interest include text mining, sentiment analysis, web mining, cyber security.

Dr. Abdulbasit A. Darem is an Associate Professor in the Department of Computer Science at Northern Border University, Saudi Arabia. He received his Ph.D. in Computer Science from the University of Mysore, India in 2014. His research interests include cyber security, malware detection, HCI, E-government, and Cloud Computing. He has published over 25 papers in top academic journals and conferences. He is a member of the IEEE. Dr. Darem is a highly accomplished researcher in the field of cyber security. His research has made significant contributions to the development of new methods for detecting and preventing malware attacks. His work has been published in top academic journals and conferences, and he has received numerous funds and awards for his research excellence

How to cite this paper: Sahana V., Anil Kumar K. M., Abdulbasit A. Darem, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on FormSpring in Textual Modality", International Journal of Computer Network and Information Security(IJCNIS), Vol.15, No.4, pp.36-47, 2023. DOI:10.5815/ijcnis.2023.04.04