TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 20, No. 1, February 2022, pp. 89~97
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v20i1.18630

Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts

Md. Tofael Ahmed[1], Maqsudur Rahman[1], Shafayet Nur[2], Abu Zafor Muhammad Touhidul Islam[3], Dipankar Das[4]
[1]Department of Information and Communication Technology, Comilla University, Comilla, Bangladesh
[2]Department of Computer Science & Engineering, Port City International University, Chattogram, Bangladesh
[3]Department of Electrical and Electronics Engineering, University of Rajshahi, Rajshahi, Bangladesh
[4]Department of Information and Communication Engineering, University of Rajshahi, Rajshahi, Bangladesh

Article Info

Article history:

Received Nov 18, 2020
Revised Dec 08, 2021
Accepted Dec 18, 2021

Keywords:

Cyberbullying
Machine learning
Natural language processing
YouTube comments

ABSTRACT

The popularity of social media has been increasing tremendously in recent times and thus cyberbullying towards people has also increased at an alarming rate. Many cyberbullying texts can be found in the comment sections of many well-known Bangladeshi social media personalities YouTube videos. It has the potential to cause severe emotional and psychological distress. Therefore, texts containing cyberbullying should be detected at the earliest stage and prevented from being displayed. In this study, we use natural language processing (NLP) techniques and various machine learning classifiers and presented model for cyberbullying detection in Bangla and Romanized Bangla texts obtained from YouTube video comments. We developed our own datasets using YouTube application programming interface (API) version 3.0. We collected 5000 Bangla comments, as well as 7000 Romanized Bangla comments from videos of different well-known social media personals. These two datasets, as well as a third dataset of 12000 texts which was the combination of the first two datasets were used to train the classifiers. These datasets were used to train machine learning classifiers after being preprocessed using NLP techniques. With an accuracy score of 76%, support vector machine (SVM) outperformed the other classifiers for the first dataset. The highest accuracy scores for the second and third datasets were 84% and 80%, respectively, which were both achieved by multinomial naive Bayes.

Corresponding Author:

Maqsudur Rahman
Department of Information and Communication Technology Comilla University
Comilla, Bangladesh
Email: mrrajon@cou.ac.bd

1. INTRODUCTION

One of the most common digital hobbies these days is spending time on social media sites such as Facebook, Instagram, Twitter, and YouTube [1]. Almost 3.6 billion people are using social media in 2020 which has an escalation rate of 49% as of January 2020. On a daily basis, people spend an average of 144 minutes on social media. This huge increase in the use of social media has its huge advantages as well as disadvantages. One of the most serious drawbacks is the rise of cyberbullying on various social media sites.

Cyberbullying is described as the intentional and repeated infliction of harm through electronic media [2]. Over 80% of children owns a mobile phone and uses social network sites of which 57% admitted the experience of cyberbullying and also 60% children and young people have witnessed bullying on social media. This horrible experience undermines a person’s freedom to use online resources and also causes several psychological effects [3]. Cyberbullying victims are 1.9 times more likely in committing suicide and furthermore endures cerebral problems like autism 75%, somatic faults 70% and learning complications 52%. This increase in cyberbullying has also escalated the necessity of prevention of cyberbullying. The detection of cyberbullying can be a significant maneuver in preventing cyberbullying. If the texts that contains cyberbullying can be detected at the earliest stage, they can be prevented from being commented.

Machine learning based classification models can be of great ply in detecting cyberbullying. Over the years machine learning models have proved their efficiency in prediction and detection. There is a huge amount of research available which utilizes machine learning based prediction and detection.

Bharat *et al.* [4] used machine learning algorithms for prediction of breast cancer and also diagnosed breast cancer using machine learning algorithms and Kaur and Kumari [5] classified diabetic and non-diabetic patients using machine learning approach. Farhana *et al.* [6] utilized deep learning approach to detect intrusion for packet and flow-based networks and in [7], again presented machine learning models for automated traffic classification and application identification. Also, Hossain *et al.* [8] used machine learning algorithms to predict rating of product reviews and in [9], the authors proposed a method of tracking and detecting vehicles from real time video streaming using blob tracker algorithm.

There are also many researches on text-based machine learning classification methods like Ikonomakis *et al.* [10] used machine learning techniques to conduct text classification. Boiy and Moens [11] used machine learning to evaluate sentiment in English, Dutch, and French texts. These kinds of text-based classification models can be a great use in classifying cyberbullying texts form regular texts. Similar kind of work was presented by Haidar *et al.* [12] where the detected cyberbullying from Arabic and English texts using machine learning models and in [13], cyberbullying from twitter of Spanish language was detected using machine learning approach. Similarly, Greevy and Smeaton [14] developed a system to detect racism using machine learning techniques. There is also research on bullying detection on Bangla texts where Al-Mamun and Akhter [15] proposed machine learning based approach.

The remaining paper is laid out as follows: section 2 includes several works that are relevant to our study. The methodology is presented in section 3. The findings are discussed in section 4, and the conclusion and future work can be found in section 5.

2. RELATED WORKS

As many researchers are working hard to detect cyberbullying in several languages, there are some previous researches available this field. In this section, we will discuss about some of the works that are relevant to our studies. Haidar *et al.* [12] proposed a solution for detecting cyberbullying using machine learning. In their research, they used both English and Romanized texts and support vector machine (SVM) had the highest overall precision (93.4%). Using SVM and naive Bayes classifiers, Dalvi *et al.* [16] proposed a machine learning model to identify and eliminate cyberbullying. The data was obtained from Twitter through the Twitter application programming interface (API). SVM had a higher accuracy of 71.25% in their analysis than naive Bayes, which had a 52.70% accuracy.

There are also several other research for cyberbullying detection, such as Paredes *et al.* [13] retrieved Spanish texts from Twitter and achieved a 93% accuracy rate using machine learning algorithms. Banerjee *et al.* [17] introduced a novel deep neural network approach for cyberbullying detection, and the CNN method received a maximum of 93.97% testing accuracy. Ali and Syed [18] also using machine learning techniques. In their research, they used three datasets and SVM had the highest average accuracy of 80%. We were inspired by these excellent efforts of cyberbullying detection in Bangla and Romanized Bangla texts.

Machine learning is also utilized in Bangla Cyberbullying detection domain. Mamun and Akhter [15] suggested using machine learning to detect cyberbullying in Bangla text. They collected 2400 status from Facebook and Twitter and applied machine learning algorithms in two phases. Their highest accuracy was 97.27% accuracy and it was gained by SVM.

Chakraborty and Seddiqui [19] used machine and deep learning to classify Bangla texts, with SVM performing best with 78% accuracy. Similarly, a maximum of 72% accuracy was achieved by Ahammed *et al.* [20]. They gathered their Bengali data from Facebook. These works for the detection of cyberbullying in Bangla motivated us to work with Bangla data collected from YouTube. Also, there is a very few works available which used Romanized Bangla texts like Tripto and Ali [21]. Their research classified sentiment of Bangla, Romanized Bangla and English texts collected from YouTube [21]. They used long short-term memory (LSTM), convolutional neural network (CNN), naive Bayes and SVM and showed an accuracy of 65% for LSTM. Similarly, Hassan *et al.* [22] used Bangla and Romanized Bangla texts. Using these texts, they trained a deep recurrent model which gave them a highest of 78% accuracy. Because the number of works for Romanized Bangla texts is minimal, we decided to conduct our research using Romanized Bangla texts collected from YouTube.

3. METHODOLOGY

3.1. Workflow

Using natural language processing techniques and machine learning classifiers, we aim to identify cyberbullying texts obtained from YouTube video comment sections. Throughout this research, a total of three datasets were used. The datasets were preprocessed using natural language processing (NLP) techniques and then were used to train the machine learning classifiers. Finally, the performance analysis was performed in terms of accuracy, precision, recall, f1-score and area under the curve of receiver characteristic operator (AUC-ROC) curve. Figure 1 depicts the proposed methodology.

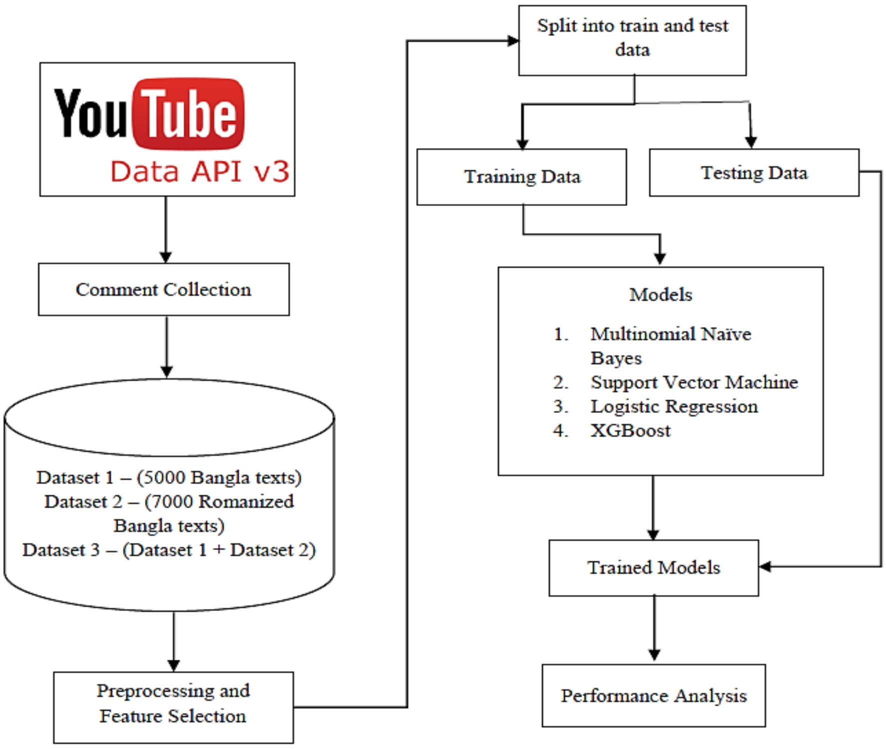


Figure 1. Proposed methodology

3.2. Dataset

The most important phase of our research is the collection of data. For this very purpose, we collected data from YouTube. For this, we utilized the YouTube API. The videos, which included a few well-known social media personalities from Bangladesh, were hand-picked. Bangla and Romanized Bangla texts were included in the texts. The texts were divided into two datasets. There were 5000 Bangla texts in Dataset 1 and 7000 Romanized Bangla texts in Dataset 2. After that, the first two datasets were combined to create a new dataset with a total of 12000 texts. Following that, we annotated all the datasets into 2 categories: bullying and non-bullying. Some of the annotated data is presented in Table 1.

Table 1. Sample of annotated data

Texts	Language	Label
ফাউল মহিলা লজ্জা করে না ছি ছি (*The foul woman is not ashamed*)	Bangla	1 (Bullying)
অরে বাল তোকে চিনিই না ....তুই বালের সিঙ্গার (*I don't even know who you are. You are a horrible singer.*)	Bangla	1 (Bullying)
সত্যিকার অর্থে ওমর সানী ভাই খুবই একটা অসাধারণ ভালো মানুষ (*Truly speaking Omor Sani brother is an awesome man.*)	Bangla	0 (Not Bullying)
ভালো লাগলো, সবাই আমার প্রিয় মানুষ, সবাই অনেক গুণবান। (*It's good, everyone is my favourite, having good quality.*)	Bangla	0 (Not Bullying)
3rd class qualityr 2 person (*Both of them are third class quality persons*)	Romanized	1 (Bullying)
Sob gula hijra magir chaoaal. (*All of them are bastards*)	Romanized	1 (Bullying)
Vlo laglo (*Feels good.*)	Romanized	0 (Not Bullying)
Bachaa digitake beshi valo lage (*I like childhood dighi even more.*)	Romanized	0 (Not Bullying)

3.3. Preprocessing

We started the preprocessing by removing any duplicate data from our datasets. All three datasets were then stripped of digits, emoticons, punctuation marks, links, user tags, uniform resource locator (URL)'s, elongated words and user mentions. Some of the texts consisted of both Bangla and Romanized Bangla which were removed from the dataset in order to obtain reliable results. Also, we did not perform any stop word removal, stemming on our datasets since the texts were mainly in local language.

3.4. Feature extraction

We used term frequency-inverse document frequency (TF-IDF) to extract features from the datasets. TF-IDF a powerful feature extraction technique which identifies important words in textual data [23]. It transforms strings into numerical values, allowing machine learning classifiers to use them. The number of times a word appears in a document divided by the total number of words in the document yields term frequency (TF).

$$TF = \frac{Frequency\ of\ a\ particular\ word\ in\ the\ document}{Total\ number\ of\ words\ in\ the\ document} \tag{1}$$

IDF identifies the weights of essential words in a document. It is measured using (2).

$$IDF = \log_2(\frac{Total\ documents}{Documents\ with\ a\ particular\ term}) \tag{2}$$

Finally, both the term frequency and inverse document frequency can be multiplied to obtain the TF-IDF which will have normalized weights. It is calculated with (3).

$$TF - IDF = TF * IDF \tag{3}$$

3.5. Machine learning classifiers

Machine learning classifiers are widely utilized to predict categorical data. Today machine learning is used to build different intelligent systems that makes decission making easier. Machine learning is a broad term that encompasses supervised, unsupervised and reinforcement learning. In this research, we used four supervised machine learning classifiers.

3.5.1. Multinomial naive Bayes

The multinomial naive Bayes algorithm is a probabilistic learning method popular in NLP. The algorithm predicts using the Bayes theorem [24]. It calculates probability for a given sample and outputs the value with the highest probability using (4).

$$P(c|x) = P(x|c) * P(c) \ / \ P(x) \tag{4}$$

3.5.2. Support vector machine (SVM)

Support vector machine (SVM) is vastly used for classification problems. It classifies data by generating a decision boundary or hyperplane in an n-dimensional space [25]. To choose the best plane among numerous possible planes, the value that has the highest margin is chosen. It has an edge over other classifiers because to its faster processing speed and greater performance with less samples.

3.5.3. Logistic regression

For binary classification, logistic regression is a commonly used classifier. To classify data, logistic regression uses a sigmoid function. The function converts any real value between 0 to 1 [26]. The sigmoid function is shown in (5).

$$S(z) = \frac{1}{1+e^{-z}} \tag{5}$$

The values that the function returns, is converted into 0 or 1. To do so, a threshold value is set. The values above the threshold value are classified as class 1 and below are classified as class 0.

3.5.4. XGBoost

XGBoost is an ensemble of decision trees [27]. It is a machine learning classifier that uses a gradient boosting algorithm. XGBoost is known for its faster execution speed and higher model performance. XGBoost is extremely useful for achieving good results with minimal resources and time.

3.6. Performance evaluation

To analyze the performance of any qualified machine learning classifier, performance evaluation is critical. We considered confusion matrix, precision, recall, f1-score, accuracy and AUC-ROC curve [28], [29] for performance evaluation. We also showed how many predictions were correctly or incorrectly done by the classifiers.

The confusion matrix is a very important performance evaluation parameter. It is a combination of for distinct actual and predicted values. Confusion matrix plays a very vital role in computing accuracy, precision, recall, f1-score and the AUC-ROC curve. The ratio of accurate predictions to the total number of input samples determines the accuracy rate [28] and is calculated using (6).

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

The number of accurate positive predictions divided by the total number of positive predictions made by a classifier yields the precision value [28] and it is calculated using (7).

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

The number of accurate positive predictions divided by the total number of actual positive samples yields the recall value [28]. It is calculated using (8).

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

The harmonic mean of precision and recall is the f1-score [28]. Better output is associated with a higher f1-score.

$$F1 - score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{9}$$

The AUC-ROC curve tells us how good a model is at distinguishing between classes [29]. A higher AUC indicates that the model is better at prediction.

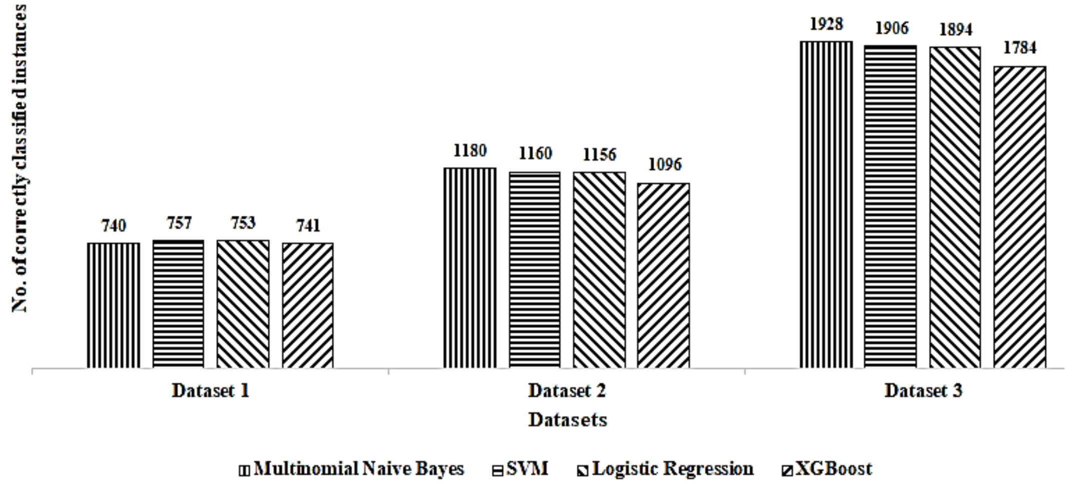
4. RESULTS AND DISCUSSION

We divided our datasets into 80% for training and 20% for testing. After training the classifiers with 80% data, we used the 20% testing sets to evaluate performance. Table 2 displays the total number of correctly and incorrectly recognized instances for each classifier across all the datasets. Figure 2 shows that SVM correctly identified 757 instances 75.7% in Dataset 1. multinomial naive Bayes has the highest number of correctly classified instances in Datasets 2 and 3, with 1180 84.28% and 1928 80.33% respectively. With the greatest number of correctly classified instances, SVM outperformed all other algorithms in Dataset 1. Similarly, multinomial Naïve Bayes stood out most for Dataset 2 and Dataset 3. Table 2 also shows the confusion matrix of the best performing algorithms for each dataset.

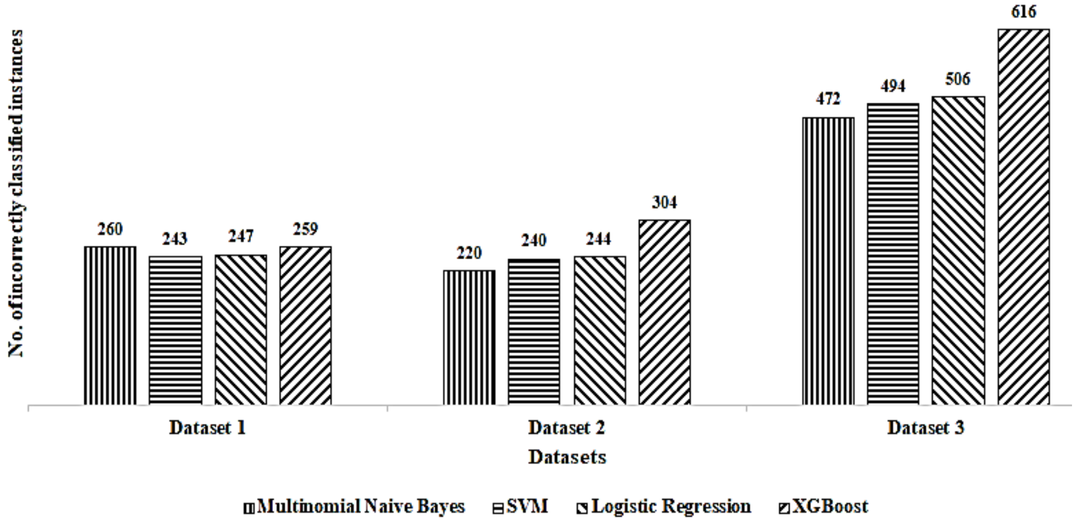
From Table 2, it can be seen that 363 cyberbullying texts and 394 non-cyberbullying texts of Dataset 1 are classified correctly by SVM for Dataset 2, 661 cyberbullying texts and 519 non-cyberbullying texts are classified correctly by multinomial naive Bayes. For Dataset 3, 1086 cyberbullying texts and 842 non-cyberbullying texts are classified correctly. Table 3 shows the precision, recall and f1-score of all algorithms for all datasets in details and Figure 3 shows the accuracy of all algorithms.

Table 3 and Figure 3 show that for Dataset 1, SVM achieved precision, recall and f1-score of 0.76 each, as well as an overall accuracy of 76%, the highest of all the algorithms. For Dataset 2, multinomial naive Bayes achieved precision, recall and f1-score of 0.84 each, as well as overall accuracy of 84%, the highest of all the algorithms. Finally, multinomial naive Bayes again outperformed all other algorithms for Dataset 3 by achieving precision of 0.81, recall and f1-score of 0.80 each, as well as an overall accuracy of 80%.

Another performance analysis is the ROC area. The larger the ROC area, the more accurately a model can identify instances. Figure 4 shows the ROC curve of SVM for Dataset 1 as well as the ROC curves of multinomial Naïve Bayes for Dataset 2 and 3 as these two algorithms performed best among all four algorithms. As shown in Figure 4, it is clear that the highest performing algorithm is multinomial naive Bayes. It performs best for Dataset 2 and 3. It also performs reasonably well for Dataset 1 but was outperformed by SVM.



(a)



(b)

Figure 2. Number of correctly and incorrectly classified instances:
(a) number of correctly classified instances and (b) number of incorrectly classified instances

Table 2. Confusion matrix of best performing algorithms for each dataset

Dataset	Algorithm	TP	FP	FN	TN
Dataset 1	SVM	363	141	102	394
Dataset 2	Multinomial Naïve Bayes	661	78	142	519
Dataset 3	Multinomial Naïve Bayes	1086	158	314	842

Table 3. Precision, recall and F1-scores by class of all algorithms for all datasets

Datasets	Algorithm	Precision	Recall	F1-score
Dataset 1	Multinomial Naïve Bayes	0.74	0.74	0.74
	SVM	0.76	0.76	0.76
	Logistic Regression	0.76	0.75	0.75
	XGBoost	0.75	0.74	0.74
Dataset 2	Multinomial Naïve Bayes	0.84	0.84	0.84
	SVM	0.83	0.83	0.83
	Logistic Regression	0.83	0.83	0.82
	XGBoost	0.80	0.78	0.78
Dataset 3	Multinomial Naïve Bayes	0.81	0.80	0.80
	SVM	0.79	0.79	0.79
	Logistic Regression	0.79	0.79	0.79
	XGBoost	0.76	0.74	0.74

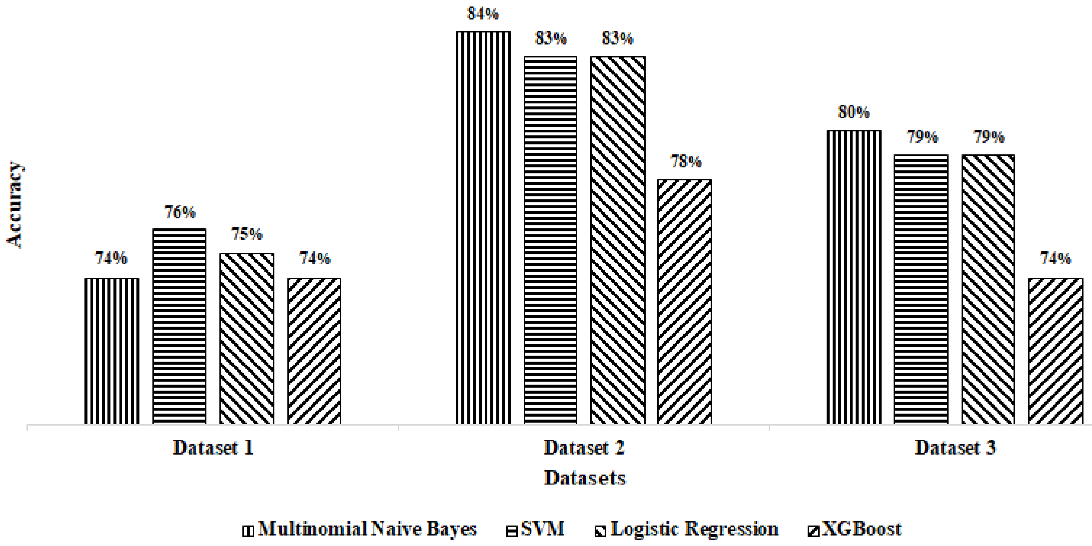


Figure 3. Accuracy of all the classifiers

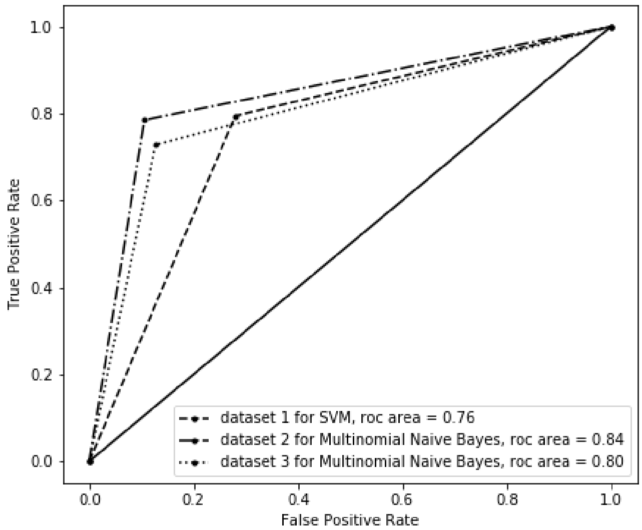


Figure 4. ROC curves for the best performing algorithms

5. CONCLUSION AND FUTURE WORK

With the increase use of different social media sites people are interacting with each other more often which has also brought an increase in the amount of cyberbullying. To detect these cyberbullying texts, we developed models based on NLP techniques and machine learning. We collected a total of 12000 texts from YouTube and prepared three datasets. These datasets were used to train the models. After testing two algorithms stood out most. SVM performed best for 1st dataset with accuracy of 76% and multinomial naive Bayes produced best results for 2nd and 3rd dataset with accuracy of 84% and 80%. SVM also obtained precision of 74%, recall 75% and f1-score 74% for first dataset. For the second dataset multinomial naive Bayes got 85% precision, 84% recall and 84% f1-score and for the third dataset multinomial Naïve Bayes got 81% precision, 80% recall and f1-scores. The other two algorithms, logistic regression and XGBoost also performed reasonably well but was slightly outperformed by SVM and XGBoost. These trained models can be used to detect Bangla and Romanized Bangla cyberbullying texts of YouTube at an early stage and stop them being commented. In the future we want to work with more data and also more videos of different categories.

REFERENCES

[1] A. Whiting and D. Williams, "Why people use social media: a uses and gratifications approach," *Qual. Mark. Res. Int. J.*, vol. 16, no. 4, pp. 362–369, 2013, doi: 10.1108/QMR-06-2013-0041.

[2] S. Hinduja and J. W. Patchin, “Cyberbullying: An exploratory analysis of factors related to offending and victimization,” *Deviant behavior*, vol. 29, no. 2, p. 129–156, 2008, doi: 10.1080/01639620701457816.
[3] A. N. Doane, M. L. Kelley, E. S. Chiang, and M. A. Padilla, “Development of the Cyberbullying Experiences Survey,” *Emerg. Adulthood*, vol. 1, no. 3, pp. 207–218, 2013, doi: 10.1177/2167696813479584.
[4] A. Bharat, N. Pooja, and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.
[5] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, 2020, doi: 10.1016/j.aci.2018.12.004.
[6] K. Farhana, M. Rahman, and M. Ahmed, "An intrusion detection system for packet and flow based networks using deep neural network approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, pp. 5514-5525, 2020, doi: 10.11591/ijece.v10i5.pp5514-5525.
[7] S. Zander, T. Nguyen and G. Armitage, "Automated traffic classification and application identification using machine learning," *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05)l*, 2005, pp. 250-257, doi: 10.1109/LCN.2005.35.
[8] M. Hossain, M. Rahman, M. Ahmed, M. Rahman, and A. Islam, "Rating Prediction of Product Reviews of Bangla Language using Machine Learning Algorithms," *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 2021, doi: 10.1109/aims52415.2021.9466022.
[9] M. Rahman, M. T. Ahmed, and A. Z. M. T. Islam, "On Detection and Tracking of the Vehicles from the Real Time Video Stream using Background Subtraction Process with Blob Tracker Algorithm," *International Journal of Ambient Systems and Applications (IJASA)*, vol. 08, pp. 21-30 10 5121, 2020, doi: 10.5121/ijasa.2020.8203.
[10] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, “Text classification using machine learning techniques,” *WSEAS Transactions on Computers*, vol. 4, no. 8, pp. 966–974, 2005.
[11] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information Retrieval*, vol. 12, no. 5, pp. 526-558, 2008, doi: 10.1007/s10791-008-9070-z.
[12] B. Haidar, M. Chamoun, and A. Serhrouchni, "A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 6, pp. 275-284, 2017. doi: 10.25046/aj020634.
[13] G. A. León-Paredes *et al.*, "Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language," *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, 2019, pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684.
[14] E. Greevy and A. Smeaton, "Classifying racist texts using a support vector machine," *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, 2004, pp. 468–469, doi: 10.1145/1008992.1009074.
[15] A.-Al-Mamun and S. Akhter, "Social media bullying detection using machine learning on Bangla text," *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, 2018, doi: 10.1109/icece.2018.8636797.
[16] R. R. Dalvi, S. B. Chavan, and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893.
[17] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 604-607, doi: 10.1109/ICACCS.2019.8728378.
[18] A. Ali and A. M. Syed, “Cyberbullying detection using machine learning,” *Pakistan Journal of Engineering and Technology*, vol. 3, no. 2, pp. 45–50, 2020, doi: 10.51846/vol3iss2pp45-50.
[19] P. Chakraborty and M. H. Seddiqui, "Threat and Abusive Language Detection on Social Media in Bengali Language," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1-6, doi: 10.1109/ICASERT.2019.8934609.
[20] S. Ahammed, M. Rahman, M. H. Niloy, and S. M. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, pp. 317-320, doi: 10.1109/SMART46866.2019.9117214.
[21] N. I. Tripto and M. E. Ali, "Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1-6, doi: 10.1109/ICBSLP.2018.8554875.
[22] A. Hassan, M. R. Amin, A. K. A. Azad, and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," *2016 International Workshop on Computational Intelligence (IWCI)*, 2016, pp. 51-56, doi: 10.1109/IWCI.2016.7860338.
[23] L. Jing, H. Huang, and H. Shi, "Improved feature selection approach TFIDF in text mining," *Proceedings. International Conference on Machine Learning and Cybernetics*. doi: 10.1109/icmlc.2002.1174522.
[24] S. Xu, Y. Li, and Z. Wang, "Bayesian Multinomial Naïve Bayes Classifier to Text Classification," *Lecture Notes in Electrical Engineering*, 2017, pp. 347-352, doi: 10.1007/978-981-10-5041-1_57.
[25] Q. Wu and D.-X. Zhou, “Analysis of support vector machine classification,” *Journal of Computational Analysis & Applications*, vol. 8, no. 2, 2006.
[26] R. E. Wright, “Logistic regression,” *American Psychological Association*, pp. 217-244, 1995.
[27] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[28] B. Juba and H. S. Le, “Precision-recall versus accuracy and the role of large data sets,” *Proc. Conf. AAAI Artif. Intell.*, vol. 33, no. 01, 2019, pp. 4039–4048, doi: 10.1609/aaai.v33i01.33014039.
[29] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

BIOGRAPHIES OF AUTHORS

Md. Tofael Ahmed Associate Professor of ICT, Comilla University (Bangladesh). Research area includes cyberbulling detection, data mining, text analytic, big data, online social network analysis. He can be contacted at email: tofael@cou.ac.bd

Maqsudur Rahman Lecturer at Department of ICT, Comilla University (Bangladesh). Teaching area includes computer programming, computer networks, data communication & mobile communication. research interests lie at digital communication, IoT and network security. He can be contacted at email: mrrajon@cou.ac.bd

Shafayet Nur Teaching Assistant at Port City International University (Bangladesh). Research interest includes machine learning, deep learning approaches, natural language processing, performance analysis. He can be contacted at email: shafayetnur.abir@gmail.com

Abu Zafor Muhammad Touhidul Islam Professor of Electrical and Electronics Engineering, Rajshahi University, Bangladesh. Completed PhD from Japan. He can be contacted at email: touhid.eee@ru.ac.bd

Dipankar Das Professor of Department of Information and Communication Engineering, University of Rajshahi, Bangladesh. Completed PhD from Japan. Research interest includes human robot interaction, computer vision. He can be contacted at email: dipankar@ru.ac.bd