An application of Machine Learning to Detect Abusive Bengali Text

Shahnoor C. Eshan
Department of Electrical and Computer Engineering
North South University, Dhaka, Bangladesh

Mohammad S. Hasan
The School of Computing and Digital Technology
Staffordshire University, Stoke-on-Trent, UK

***Abstract*—Bengali abusive text detection can be useful to prevent cyberbullying and online harassment as these types of crimes are increasing rapidly in Bangladesh. Machine learning approach can be useful to keep the system always updated with the new types of approaches used by the abusers. This paper investigates machine learning algorithms e.g. Random Forest, Multinomial Naïve Bayes, Support Vector Machine (SVM) with Linear, Radial Basis Function (RBF), Polynomial and Sigmoid kernel and have compared with unigram, bigram and trigram based CountVectorizer and TfidfVectorizer features. The results show that SVM Linear kernel performs the best with trigram TfidfVectorizer features.**

Keywords—Random Forest, Multinomial Naïve Bayes, Neural Network, Support Vector Machine etc.

I. Introduction

Social networks are becoming very integral part of human life. Facebook is the most popular social network which has 2.01 billion monthly active users [1]. Instagram is one of the most popular photo sharing platform with more than 700 million users [2]. YouTube is the most popular video sharing platform with 1.5 million users [3]. Day by day sharing information is getting easier; however, cyberbullying is also getting more common. Paper [4] explained the negative impacts of cyberbullying on children and another paper [5] shows that victims have more suicidal thoughts than non-victims. A 19 years old girl committed suicide because of hateful messages she received from two boys [6]. Cybercrime and cyberbullying are on the rise in Bangladesh [7], [8]. About 73 percent women Internet users in Bangladesh are in a threat of online harassment [9]. These types of incidents mostly happen through comments, messages or via negative images.

There have been several types of research on text classification, object detection etc. Methods used in those researches can be used to detect abusive messages, comments or vulgar images. As the online abusers change their way of action, Machine Learning (ML) can be a very useful method to learn their behaviour. This can play a significant role to find Internet predators and eliminate their actions. This paper investigates the application of ML algorithms to detect abusive texts and evaluates performances. The training and testing samples are collected from the comments on posts of Bangladeshi Facebook celebrities so that the performances of the ML algorithms can be understood with real-world samples.

The remaining of the paper is organized as follows. Section II shows the previous works in the similar field. Section III explains the ML algorithms that have been used for the experiments. Section IV has details about the libraries and tools used for the experiments, explanation about the text features being used for the experiments and how the results are being validated. Section 0 contains results obtained from the conducted experiments and analysis. Finally, the paper concludes in section VI.

II. Existing works and Limitations

Many researches have been conducted on text classification and its applications. In paper [10], Support Vector Machine (SVM), C4.5, and Naïve Bayes (NB) algorithms are compared for text classification where SVM outperforms. The results of the research are shown in Table 1 and Table 2. Paper [11] compares text classification performance on different supervised ML models where back-propagation based neural network has the best performance. The results are shown in Figure 1.

Table 1: Diabetes Dataset Results [10].

% Split of training set	Algorithm		
	NB	C4.5	SVM
At 66% (261 instances)	77.01	76.24	79.31
At 90% (77 instances)	77.9	75.32	80.52
At 33% (515 instances)	73.98	70.29	75.73
Precision (Weighted Avg.)	0.767	0.756	0.787
Recall (Weighted Avg.)	0.77	0.762	0.793

Table 2: Calories Dataset Results [10].

% Split of training set	Algorithm		
	NB	C4.5	SVM
At 66% (14 instances)	78.57	78.57	71.52
At 90% (4 instances)	100	75	75
At 33% (27 instances)	81.48	85.18	66.67
Precision (Weighted Avg.)	0.844	0.802	0.81
Recall (Weighted Avg.)	0.786	0.786	0.714

Sentiment analysis is one of the common text classification problems and several types of researches have been conducted on sentiment analysis on Bengali text. MaxEnt and SVM algorithms are compared in paper [12] for sentiment analysis on Bengali microblog posts with different feature extraction methods and it gets the best performance with SVM with unigram and emoticons as features. Also, sentiment analysis is

done on Bengali horoscope corpus in paper [13] using ML algorithms e.g. NB, SVM, K-Nearest Neighbors (NN), Decision Trees (DT), and Random Forest (RF). Among those SVM has the best performance with unigram features and Figure 2 shows the results. In paper [14], a survey is conducted on text content filtration using different text categorization methods. Paper [15] has shown an implementation of web content filtration by using string searching algorithm Aho-Corasick.

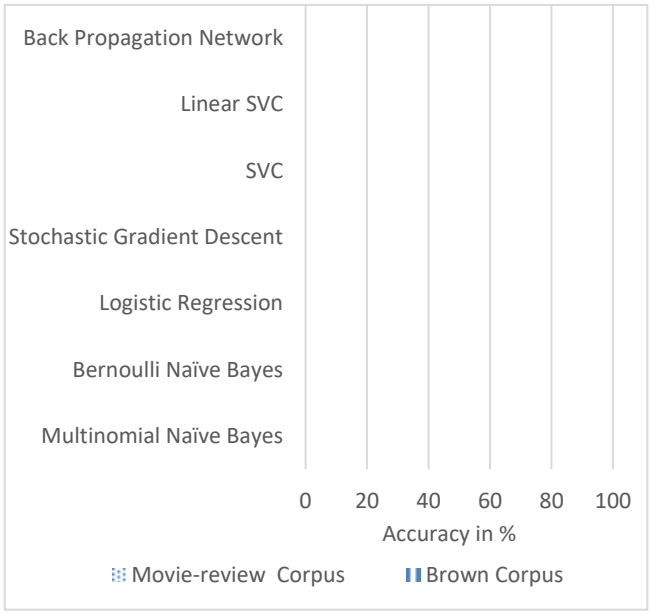

Figure 1: Result of accuracy (%) of text classification [11].

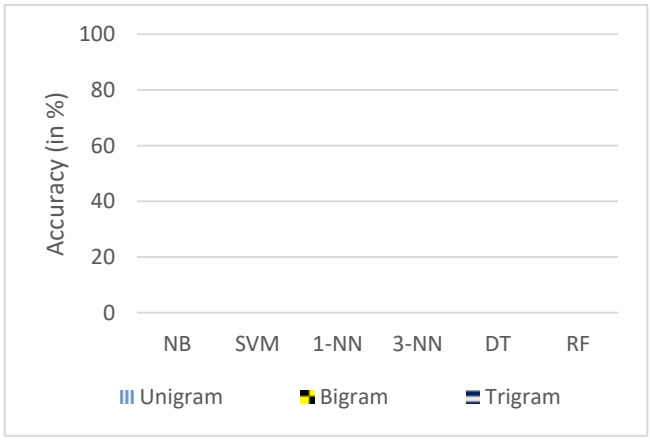

Figure 2: 10-fold cross-validation results considering all features [13].

Although there are researches on text classification, sentiment analysis on Bengali text and offensive text filtration, no or limited experiment has been conducted on ML-based Bengali abusive text detection. The Bengali abusive text dataset used in this paper is collected from Facebook comments where sentences can be grammatically incorrect and can have spellings mistakes and it is completely different from the other types of datasets. There is no analysis on how ML algorithms may perform with this type of data, what types of text feature can provide more accuracy with ML algorithms, these issues are being considered in this paper.

III. Basics of Machine Learning Algorithms

A. *Multinomial Naïve Bayes (MNB)*

NB classifiers are probabilistic classifiers based on Bayes theorem, and MNB is a variation of it. It considers the number of repetitions of a team in training dataset. MNB performs better than Bernoulli Naïve Bayes (BNB) for text classification [16].

B. *Random Forest (RF)*

RF creates multiple random decision trees during training time. During prediction, it collects results from each tree and returns the result which has been predicted by most of the trees. The RF classifiers used for the experiments in this paper has 10 trees in the forest.

C. *Support Vector Machine (SVM)*

Support Vector Machine is a very popular ML algorithm and from training data it returns an optimal hyperplane to categorize new samples.

SVM classifiers use different kernels to build an optimal hyperplane. Kernels are functions which analyze the patterns. The following SVM kernels are used for the experiments in this paper.

1) Linear – Linear kernel often performs better when the dataset is linearly separable. The kernel function is $\langle x, x' \rangle$ [17].
2) Radial Bias Function (RBF) – It represents two samples x and x (prime) as feature vectors in an input space defined as ($\gamma\langle x, x' \rangle$+r)d – where d is degree and r is coef0 [17].
3) Polynomial – Polynomial kernel allows learning of nonlinear samples. The kernel function is defined as exp(-γ||x - x′||2) . γ – where gamma (γ) must be greater than 0 [17].
4) Sigmoid – The SVM Sigmoid kernel function is tanh($\gamma\langle x, x' \rangle$+r) – where r is specified by coef0 [17].

IV. Experiment Design

For the experiments conducted in this paper comments are scrapped from posts on several Facebook celebrities in Bangladesh e.g. Shakib Al Hasan [18], Naila Nayem [19], Rasmi Alon [20], Mehenil Tasnim Joya [21], TunTuni AdRita [22], Hero Alom [23], Mahmudullah Riyad [24], Najnin Akter Happy [25], Ananta CIP [26], Barsha [27], Tahsan [28], Nusraat Faria [29], Sabila Nur [30], Shakib Khan [31], Nusrat Imrose Tisha [32], Bidya Sinha Saha Mim [33], Asif Akbar [34], RJ Tazz [35] and Model Arif Khan [36]. To ensure privacy, only public comments are scrapped and commenters' name and id are not being scrapped. After scrapping comments, all the special characters e.g. '(', '-', '@' etc., Unicode emoticons are removed. In this paper, the comments containing only Bengali Unicode characters are considered. Then each Bengali Unicode comment is then categorized either as abusive or non-abusive manually.

To validate the results, 10 times 10 folds cross-validation method is used. It means all the datasets are folded 10 times where each time 90% data is used for training and 10% data is

used for validation and after 10 folds the average is counted. This process is being run 10 times and the average is counted. Tests are being done with 500, 1000, 1500, 2000 and 2500 comments scrapped from Facebook. For each test, random comments are loaded from the database where 50% are abusive and rest is non-abusive.

Table 3: Sample sentences and categories.

Sentence	Category (cat)
man and woman are helpless without each other	A
woman and man are helpless without each other	B
without each other man and woman are helpless	C
without each other woman and man are helpless	D

To find out which algorithms perform better with what type of features, experiments are conducted with three types of string features which are unigram, bigram, and trigram. There could be several examples in training or testing sets which uses almost same features in different order, in those cases, bigram and trigram might be able to give better results than unigram. If the sentences given in Table 3 are classified differently, unigram may not perform properly.

A. Unigram

Unigram features do not consider the relations between the words. Each word in the sentence is a feature. So, for the dataset given in Table 3, the features are given in Table 4.

Table 4: Unigram features for the dataset given in Table 3.

Serial #	Feature	Serial #	Feature
1.	and	2.	are
3.	each	4.	helpless
5.	man	6.	other
7.	without	8.	woman

Table 5 shows the vector representation of the used dataset from the unigram features above.

Table 5: Unigram features vector representation.

cat	Features							
	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1

B. Bigram

Bigram features consider the relation between two consecutive words. So, for the dataset given in Table 3, the trigram features are given in Table 8.

Table 6: Bigram features for the dataset given in Table 3.

Serial #	Feature	Serial #	Feature
1.	and man	2.	and woman
3.	are helpless	4.	each other
5.	helpless without	6.	man and
7.	man are	8.	other man
9.	other woman	10.	without each
11.	woman and	12.	woman are

Table 7 shows the vector representation of the Table 3 dataset from the bigram features in Table 6.

Table 7: Bigram features vector representation.

cat	Features											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	1	1	1	1	0	0	0	1	0	1
2	1	0	1	1	1	0	1	0	0	1	1	0
3	0	1	1	1	0	1	0	1	0	1	0	1
4	1	0	1	1	0	0	1	0	1	1	1	0

C. Trigram

Trigram features consider the relationship between three consecutive words in a sentence. So, for the dataset given in Table 3, the trigram features are given in Table 8. The vector representation of the trigram features in Table 8 from the considered dataset in Table 3 is shown in Table 9.

To train the ML algorithms, these features are then used for vector creation. For the experiments in this paper, two types of vectors are used – CountVectorizer [37] and TfidfVectorizer [38]. CountVectorizer takes into account the frequency of features. On the other hand, TfidfVectorizer tries to determine the importance of a feature so that the classifier does not miss less frequent but important features. The hardware, tools etc. are explained in Table 10.

Table 8: Trigram features for the dataset given in Table 3.

Serial #	Feature	Serial #	Feature
1.	and man are	2.	and woman are
3.	are helpless without	4.	each other man
5.	each other woman	6.	helpless without each
7.	man and woman	8.	man are helpless
9.	other man and	10.	other woman and
11.	without each other	12.	woman and man
13.	woman are helpless		

Table 9: Trigram features vector representation.

cat	Features												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	0	0	1	1	0	0	0	1	0	1
2	1	0	1	0	0	1	0	1	0	0	1	1	0
3	0	1	0	1	0	0	1	0	1	0	1	0	1
4	1	0	0	0	1	0	0	1	0	1	1	1	0

Table 10: Hardware and tools

Hardware or Tool	Description
Processor	Intel Core i7 6700HQ (6MB cache, 2.60 GHz)
Memory	8GB 2100MHz DDR4
Hard disk drive	500GB 7100RPM HDD
Operating system	Ubuntu 16.04
Programming Language	Python
Numpy	For larger array handling
Scikit-Learn	Feature extraction and ML algorithm implementation
Django	For handling commands, scrapper, ORM for database handling and admin system developed for categorizing scrapped comments

V. RESULTS

The unigram, bigram and trigram features are extracted from all the comments and vectorized using CountVectorizer and TfidfVectorizer to train the machine learning algorithms. The results are then validated using 10 times 10 folds cross-validation method and are explained in the following sections.

A. *CountVectorizer vector*

Figure 3 shows the results acquired from the ML algorithms with unigram features and CountVectorizer. It shows that SVM with Linear kernel performs the best in terms of accuracy while SVM with Sigmoid kernel has the poorest performance. Performance improves in RF, MNB and SVM (Linear) as the number of training data goes higher.

In Figure 4, the results of the ML algorithms with Bigram CountVectorizer features are shown. MNB shows the best accuracy and SVM with Linear kernel is very close to it. Performances of all the algorithms increase with the number of datasets except SVM Polynomial kernel which remains almost the same.

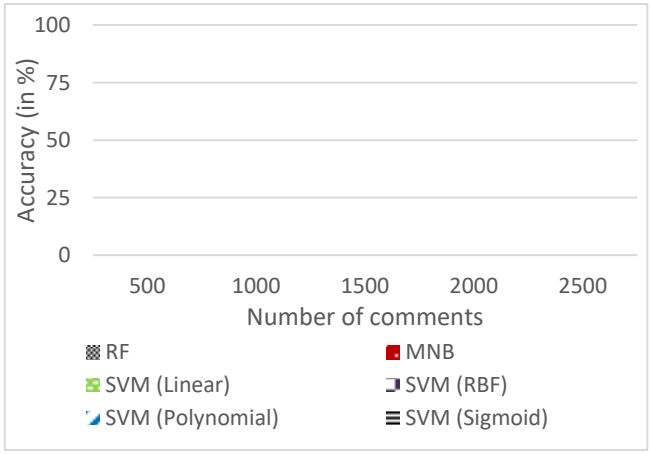

Figure 3: Unigram with CountVectorizer.

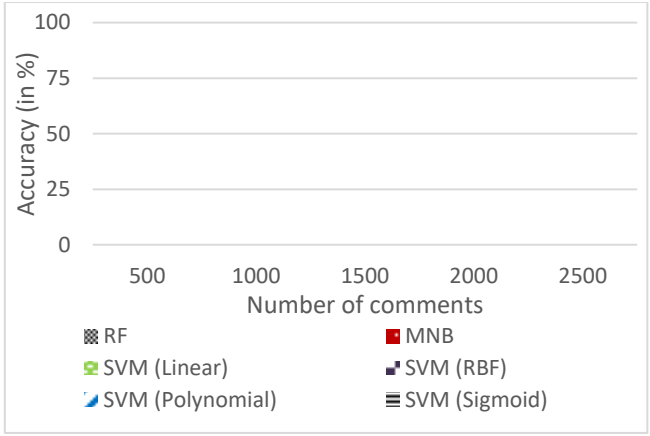

Figure 4: Bigram with CountVectorizer.

In Figure 5, RF, MNB, and SVM with Linear Kernel show much better accuracies than those of SVM with RBF, Polynomial and Sigmoid for Trigram CountVectorizer. The performance gap between those two groups is quite noticeable. MNB has the best performance which is the highest accuracy received from these algorithms among all the experiments. This is the highest accuracy of RF among all the experiments conducted in this paper. Performance of SVM RBF and Sigmoid kernel drops with the number of datasets while Polynomial kernel is not showing any change at all.

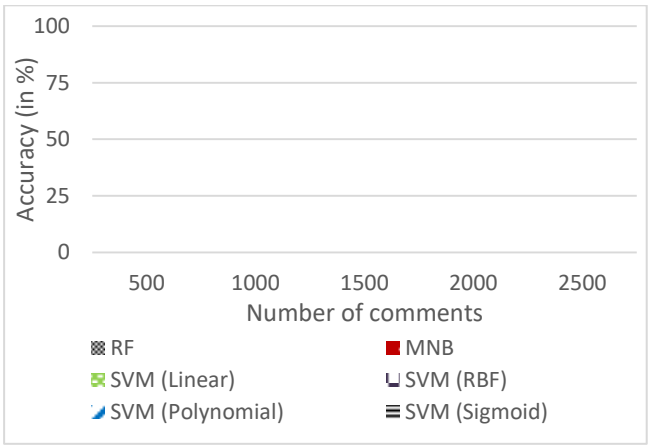

Figure 5: Trigram with CountVectorizer.

B. *TfidfVectorizer vector*

The ML algorithms are trained with unigram TfidfVectorizer and the results are shown in Figure 6 for Unigram. SVM with Linear kernel shows the best performance. Although MNB has very good accuracy in some previous experiments, in this experiment it results poor. Also, the accuracy of SVM with RBF kernel is very close to the SVM with the Sigmoid kernel which is the poorest in terms of accuracy.

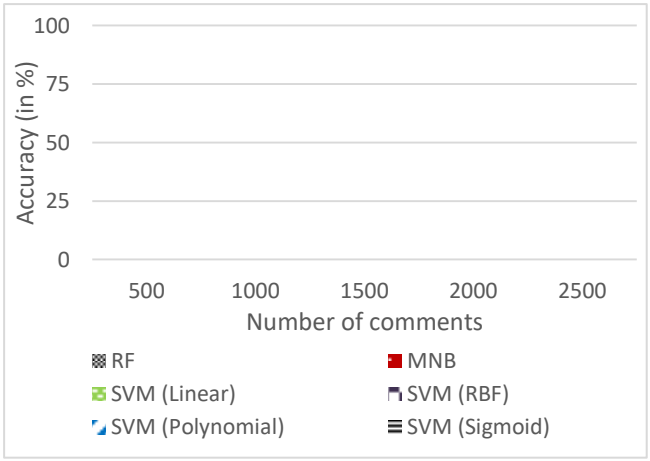

Figure 6: Unigram with TfidfVectorizer.

Figure 7 shows the performances of the algorithms for Bigram TfidfVectorizer features. SVM with Linear Kernel has the best accuracy. SVM with RBF, Polynomial and Sigmoid are in last three positions.

Figure 8 shows the result received from the experiments conducted with Trigram TfidfVectorizer. Among all the experiments the best result received from this experiment with SVM Linear kernel classifier. MNB and RF are in second and third places in terms of accuracy.

From the experiments conducted, it is noted that SVM with Linear kernel performs the best in most cases. The dataset might be the reason e.g. probably the dataset is linearly separable. On the other hand, RBF, Polynomial and Sigmoid kernels of SVM cannot perform significantly well. The probable reason behind the poor performance of those kernels could be the decision boundary which they are considering are not optimal enough to fit different classes of text training dataset properly. MNB has very good accuracies in most cases; it gets better with bigram and trigram models, which indicates that it finds more similar features in those cases. Although RF generates 10 random decision trees which might not be the most optimally distributed, on average those trees provide accurate results and that is the reason behind the good performance of RF.

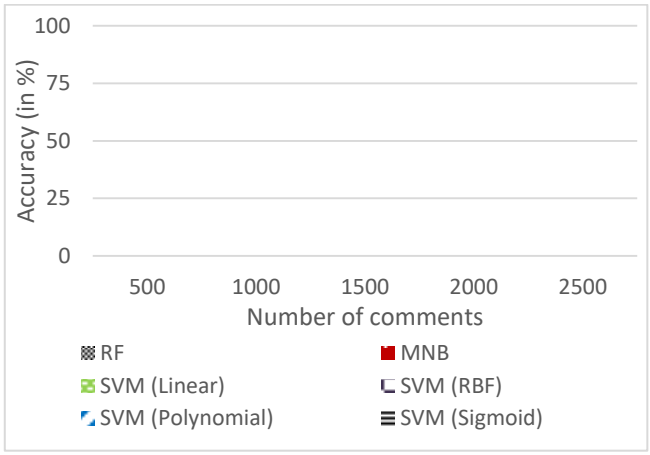

Figure 7: Bigram with TfidfVectorizer.

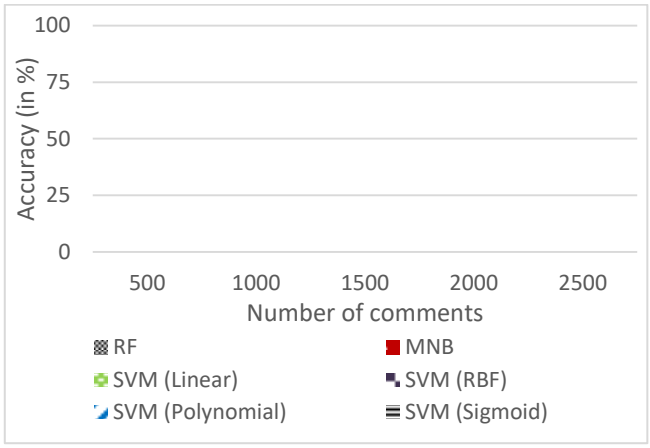

Figure 8: Trigram with TfidfVectorizer.

SVM Linear kernel provides the highest accuracy among other algorithms with trigram TfidfVectorizer. Trigram considers three consecutive words as a feature, which makes the pattern of the features more distinguishable. On the other hand, tf-idf may increase the weight of less frequent but important feature. This might be the reason behind the performance achieved from this combination. Even MNB shows significant accuracy when trained with trigram TfidfVectorizer. MNB provides the highest accuracy with trigram CountVectorizer, which indicates that trigram features are more distinguishable, and probably significance of features does not vary much with trigram, that could be the reason why MNB performed the best with trigram feature frequency dataset.

VI. Conclusion

From the experiments, it is clearly visible that among all the ML algorithms used in this paper, TfidfVectorizer features with SVM linear kernel performs the best when compared with CountVectorizer features. Also, MNB can be considered for abusive text classification as it performs well in most cases. As future works, experiments may be conducted with neural network based models like Deep Neural Network (DNN), Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN).

Acknowledgment

The authors would like to acknowledge an Erasmus+ International Credit Mobility (ICM) fund for Bangladesh awarded to Dr. Mohammad Hasan at Staffordshire University, UK in 2016.

References

[1] "Company Info | Facebook Newsroom." [Online]. Available: https://newsroom.fb.com/company-info/#statistics. [Accessed: 06-Aug-2017].

[2] "Instagram Blog." [Online]. Available: http://blog.instagram.com/post/160011713372/170426-700million. [Accessed: 06-Aug-2017].

[3] "YouTube has 1.5 billion logged-in monthly users watching a ton of mobile video | TechCrunch." [Online]. Available: https://techcrunch.com/2017/06/22/youtube-has-1-5-billion-logged-in-monthly-users-watching-a-ton-of-mobile-video/. [Accessed: 06-Aug-2017].

[4] C. Nixon, "Current perspectives: the impact of cyberbullying on adolescent health," *Adolesc. Health. Med. Ther.*, p. 143, 2014.

[5] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, 2010.

[6] "Teenage girl tragically killed herself after bullies flooded her social media accounts with horrible messages... as her parents call for action against the two boys responsible | Daily Mail Online." [Online]. Available: http://www.dailymail.co.uk/news/article-3043651/Parents-teen-committed-suicide-suffering-vicious-bullying-Facebook-demand-greater-protection-kids-online.html. [Accessed: 07-Aug-2017].

[7] "Cybercrime cases on the rise." [Online]. Available: http://en.prothom-alo.com/bangladesh/news/122235/Cybercrimes-on-the-rise-due-to-section-57. [Accessed: 07-Aug-2017].

[8] "Digital Sexual Harassment in Digital Bangladesh | The Daily Star." [Online]. Available: http://www.thedailystar.net/in-focus/digital-sexual-harassment-digital-bangladesh-82480. [Accessed: 07-Aug-2017].

[9] "73 percent women subject to cyber-crime in Bangladesh - bdnews24.com." [Online]. Available: http://bdnews24.com/bangladesh/2017/03/09/73-percent-women-subject-to-cyber-crime-in-bangladesh. [Accessed: 07-Aug-2017].

[10] M. Trivedi, N. Soni, S. Sharma, and S. Nair, "Comparison of Text Classification Algorithms," vol. 4, no. 2, pp. 334–336, 2015.

[11] S. Z. Mishu and S. M. Rafiuddin, "Performance Analysis of Supervised Machine Learning Algorithms for Text Classification," *2016 19th Int. Conf. Comput. Inf. Technol.*, pp. 409–413, 2016.

[12] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," *2014 Int. Conf. Informatics, Electron. Vision, ICIEV 2014*, 2014.

[13] T. Ghosal, S. K. Das, and S. Bhattacharjee, "Sentiment analysis on (Bengali horoscope) corpus," *12th IEEE Int. Conf. Electron. Energy, Environ. Commun. Comput. Control (E3-C3), INDICON 2015*, pp. 1–6, 2016.

[14] S. H. Yadav and B. L. Parne, "A survey on different text categorization techniques for text filtration," *Proc. 2015 IEEE 9th Int. Conf. Intell. Syst. Control. ISCO 2015*, 2015.
[15] S. H. Yadav and P. M. Manwatkar, "An approach for offensive text detection and prevention in Social Networks," *ICIIECS 2015 - 2015 IEEE Int. Conf. Innov. Information, Embed. Commun. Syst.*, pp. 3–6, 2015.
[16] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *AAAI/ICML-98 Work. Learn. Text Categ.*, pp. 41–48, 1998.
[17] "1.4. Support Vector Machines — scikit-learn 0.19.0 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/svm.html#kernel-functions. [Accessed: 30-Aug-2017].
[18] "Shakib Al Hasan - Home | Facebook." [Online]. Available: https://www.facebook.com/Shakib.Al.Hasan/. [Accessed: 13-Aug-2017].
[19] "Naila Nayem - Home | Facebook." [Online]. Available: https://www.facebook.com/artist.nailanayem/. [Accessed: 13-Aug-2017].
[20] "Rasmi Alon - Home | Facebook." [Online]. Available: https://www.facebook.com/komolrasmi1/. [Accessed: 13-Aug-2017].
[21] "Mehenil Tasnim Joya - Home | Facebook." [Online]. Available: https://www.facebook.com/khelahoba. [Accessed: 13-Aug-2017].
[22] "TunTuni AdRita - Home | Facebook." [Online]. Available: https://www.facebook.com/official.page.TunTuni. [Accessed: 13-Aug-2017].
[23] "Hero Alom - Home | Facebook." [Online]. Available: https://www.facebook.com/alomofficial. [Accessed: 13-Aug-2017].
[24] "Mahmudullah Riyad - Home | Facebook." [Online]. Available: https://www.facebook.com/Mahmudullah.Riyad. [Accessed: 13-Aug-2017].
[25] "Najnin Akter Happy - Home | Facebook." [Online]. Available: https://www.facebook.com/NajninAkterBD. [Accessed: 13-Aug-2017].
[26] "Ananta CIP - Home | Facebook." [Online]. Available: https://www.facebook.com/AnantaOfficial. [Accessed: 13-Aug-2017].
[27] "Barsha - Home | Facebook." [Online]. Available: https://www.facebook.com/BarshaOnline. [Accessed: 13-Aug-2017].
[28] "Tahsan - Home | Facebook." [Online]. Available: https://www.facebook.com/tahsanfans. [Accessed: 13-Aug-2017].
[29] "Nusraat Faria - Home | Facebook." [Online]. Available: https://www.facebook.com/nusraatfariaofficial. [Accessed: 13-Aug-2017].
[30] "Sabila Nur - Home | Facebook." [Online]. Available: https://www.facebook.com/Sabilanursablabablaofficial. [Accessed: 13-Aug-2017].
[31] "Shakib Khan - Home | Facebook." [Online]. Available: https://www.facebook.com/Iamshakibkhanbd. [Accessed: 13-Aug-2017].
[32] "Nusrat Imrose Tisha - Home | Facebook." [Online]. Available: https://www.facebook.com/TishaBDactressOfficial. [Accessed: 13-Aug-2017].
[33] "Bidya Sinha Saha Mim - Home | Facebook." [Online]. Available: https://www.facebook.com/BidyaSinhaSahaMim. [Accessed: 13-Aug-2017].
[34] "ASIF - Home | Facebook." [Online]. Available: https://www.facebook.com/asif.akbar.bd. [Accessed: 13-Aug-2017].
[35] "Rj Tazz - Home | Facebook." [Online]. Available: https://www.facebook.com/RjTazzOfficial. [Accessed: 13-Aug-2017].
[36] "Model Arif Khan - Home | Facebook." [Online]. Available: https://www.facebook.com/modelarifkhan. [Accessed: 13-Aug-2017].
[37] "sklearn.feature_extraction.text.CountVectorizer — scikit-learn 0.19.0 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html. [Accessed: 13-Aug-2017].
[38] "sklearn.feature_extraction.text.TfidfVectorizer — scikit-learn 0.19.0 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. [Accessed: 24-Aug-2017].