Social media bullying detection using machine learning on Bangla text

Abdhullah-Al-Mamun[1,] Shahin Akhter[1]

[1]IICT, BUET
Dhaka, Bangladesh
mamun_kuet04@yahoo.com

Abstract—**with the popularity of Unicode system and growing use of Internet, the use of Bangla over social media is increasing. However, very few works have been done on Bangla text for social media activity monitoring due to a lack of a large number of annotated corpora, named dictionaries and morphological analyzer, which demands in-depth analysis on Bangladesh's perspective. Moreover, solving the issue by applying available techniques is very content specific, which means that false detection can occur if contents changed from formal English to verbal abuse or sarcasm. Also, performance may vary due to linguistic differences between English and non-English contents and the socio-emotional behaviour of the study population. To combat such issues, this paper proposes the use of machine learning algorithms and the inclusion of user information for cyber bullying detection on Bangla text. For this purpose, a set of Bangla text has been collected from available social media platforms and labelled as either bullied or not bullied for training different machine learning based classification models. Cross-validation results of the models indicate that a support vector machine based algorithm achieves superior performance on Bangla text with a detection accuracy of 97%. Besides, the impact of user specific information such as location, age and gender can further improve the classification accuracy of Bangla cyber bullying detection system.**

***Index Terms*—Cyber bullying, Bangla text, Classification, Machine learning**

I. Introduction

Cyber bullying is when someone uses technology to send threatening or embarrassing messages to another person. Bullying on social media can be even worse due to its quick spread to the wider audience. Research shows that such behaviour frequently occurred on Facebook and Twitter sites [1]. Out of 80.83 million Internet users in Bangladesh [2], more than 90% of social media users are active on Facebook where majority is young, vulnerable and in dare need of protection. With being 7th most spoken language and with the popularity of Unicode system and growing use of Internet [3], the use of Bangla over social media is increasing. But very few works have been done on Bangla text for social media activity monitoring due to a lack of a large number of annotated corpora, named dictionaries and morphological analyzer [3], which demands in-depth analysis on Bangladesh's perspective. Due to its quick spread to the wider audience, Cyber bullying has been receiving a profound interest to the researchers over the past few years. However, Bangla, the seventh most widely spoken language in the world and the second most used language in the Indian subcontinent, is lagging far behind in Bangla cyber bullying research. Meanwhile, with progressive affordability of technology and incentives by the concerned Governments, Bangla document analysis is more relevant now than ever before.

In general, the data available on social media are mostly short, noisy, unstructured and sometimes a mix-up of multiple languages, which means that the use of traditional bullying detection methods like guidelines, human moderations, and keyword searches will fall short in social media data [4]. Research shows use of machine learning [5] algorithms and sentiment analysis [4] for social media data has better accuracy than the keyword search and textual analysis of contents [6-7]. But proposed machine learning technique in literature are very content specific. Due to linguistic differences between English and non-English contents, performance may vary. Besides, the socio-emotional behaviour and user specific information of the study population also has significant impact for cyber bullying detection. For instance, support vector machine (SVM), a popular learning method for English text was less accurate on Arabic texts [11] compared to NB. Hence, this project intends to investigate cyber bullying further to apply on Bangla text. The objective of this work is to develop a cyber bullying detection and monitoring scheme suitable for Bangla text on social media network. Hence, targets of this paper are:

1. To design a novel scheme for analysing Bangla content on social media by combining text analytics and machine learning algorithms.
2. To compare the performance of the module with other available techniques.

The rest of the paper is organized as follows. In section II, existing cyber bullying detection methods are discussed. Section III provides the motivation of the current approach and methodologies used for developing the cyber bullying detection scheme for Bangla text. Section IV presents the classification outcome of the proposed method and section V contains the concluding remarks of the overall method.

II. Background

For English language, a notable amount of research have been performed in text categorization or cyber bullying detection.[1,13] using text mining by classifying posts or conversations. Yin, Xue and Hong [7] applied supervised learning for text classification by labelling texts using N-grams technique and weighting using term frequency (TF)-inverse document frequency (IDF). Dinakar, Reichart and Lieberman [5] conducted a comparison on text classification techniques using various supervised approach. The research [5] included YouTube comments, each was manually labelled and then various binary and multiclass classifications were implemented. Decision tree (J48) and k-nearest neighbour (KNN) were used by Kelly Reynolds [6]. Among the available classification techniques, SVM gets notable attention due to better performance in various text classifications. In [17],

978-1-5386-7482-6/18/$31.00 ©2018 IEEE

author argued the theoretical perspective of applying SVM on text classification. Zhijie et al. [18] compared SVM against naive bayes (NB) and k-nearest neighbour (KNN), and observed that SVM outperformed the others. But all of the aforementioned research was performed on only on English text.

Due to linguistic differences between English and non-English contents, performance and accuracy of algorithm varies on non-English content. A recent study reported that the NB classifier can be used effectively for Indian text classification [12] problem. In [19], a combination of NB and Ontology based classification showed better performance on Punjabi text. In [12], researcher showed that classification results of SVM was better than the NB method for Urdu language. Also, Artificial Neural network model based method showed better performance than Vector Space Model for Tamil contents [20]. Decision Tree, Neural Networks, and N-gram are also potential techniques in text classification for non-English content. Very few works has been done on bangle text. Hence, the aim of this research was to explore various machine learning algorithms on Bangla text and identify the best performing one for detection of cyber bullying on social media contents.

III. Experimental Setup

A. *Proposed Model*

The proposed algorithm consists of four main components as shown in Fig. 1.

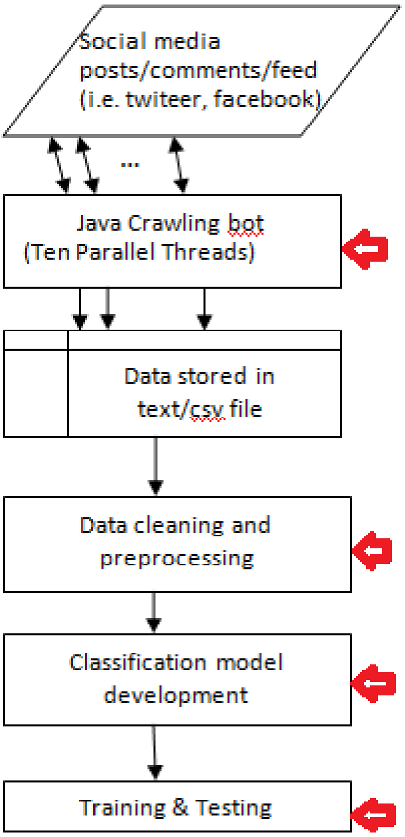

Fig. 1: A graphical representation of the proposed social media bullying detection algorithm.

The main function of first component, java crawler, was to fetch/extract social media public data using available API (i.e. Twitter API, Facebook graph API etc). Ten parallel threads were developed to fetch the Bangla social media content on 30 minute interval and save the data into text/csv file. The second module was for data cleaning and pre-processing. As the extracted data had multi-lingual unstructured content along with lot of emoji (ideograms and smileys used in electronic messages and web pages), it was required to clean the data for higher accuracy. The third component of this setup was the machine learning model. Several supervised machine learning algorithms were considered for training and testing and performance were compared to identify the best one. Besides user specific features i.e. demographic, geographic, time related etc. data were also included as potential features for cyber bullying detection.

B. *Data sets*

A java program was developed for extracting social media data. Facebook and Twitter sites were considered for Bangla text content collection. Using Facebook Graph API, a total of 1000 contents were collected including Bangla public status, emotion icons and user related information. Besides, 1400 Bangla public status, user's demographic and geographic data were fetched from Twitter using Twitter REST API. All of these collected texts were manually labelled as either bullied or not-bullied. The bag of word for technique was incorporated to facilitate the training model development.

C. *Data pre-processing and feature extraction*

Extracted Bangla text content was pre-processed and tokenized by separating special character and emoji from text data. Text based features such as TF-IDF and user information related to the online activity and connectivity network were separated for later usage. Stemming and tokenization has been applied on text contents to facilitate the feature extraction. Porter Stemmer, removing the commoner morphological and inflexional endings, was included considering its wide scale application in non-English contents [19]. Next, a tri-gram model and word tokenization was applied for text based feature extraction.

D. *Classification model development*

An exhaustive exploration has been performed to identify suitable machine learning algorithm for Bangla text categorization. Available machine learning models from literature has been considered (i.e. NB, SVM, Decision Tree and KNN). WEKA software platform has been used for this purpose. The experiments included 2,400 Bangla texts collected from social media posts where of the data contained 10% bullying text. A 10-Fold cross validation model was followed for testing the performance of the developed models. Overall, the training and testing process was carried in two phases: i) including only text based features (posts/comments) ii) including both text based features and user information. The following 11 features related to the user information were collected from social media:

1. Public post/comments
2. Favourite count
3. Post time
4. Re-Tweet count
5. User status count
6. User public description
7. Follower count
8. Friends count
9. Location
10. Screen name

978-1-5386-7482-6/18/$31.00 ©2018 IEEE

11. Gender

IV. Results and Discussion

Table I presents the outcome of the classification algorithms from the 1st phase of experiment. Results indicate that SVM based classification model performs best on Bangla contents from social media. The method achieved 95.40% accuracy after 10-fold cross-validation by separating bullied text contents from the non-bullied ones.

TABLE I
Performance of Social MEDIA BULLYING detection ON BANGLA using user's posts only

Algorithm	Precision	F1-Score	ROC	Accuracy
Naïve Bayes	0.87	0.69	0.67	60.98%
J48	0.82	0.86	0.50	90.57%
SVM	0.95	0.94	0.76	95.40%
KNN (1-Nearest)	0.92	0.90	0.61	92.50%
KNN (3-Nearest)	0.91	0.86	0.75	90.81%

In the 2nd phase of the experiment, both text based features and user information has been included in training the NB, SVM, J48 and KNN classifier models. The result of this experiment has been presented in Table II.

TABLE III
performance of Social MEDIA BULLYING detection ON BANGLA using posts and user specific data

Algorithm	Precision	F1-Score	ROC	Accuracy
Naïve Bayes	0.98	0.66	0.90	50.73%
J48	0.99	0.99	0.98	97.23%
SVM	0.99	0.99	0.71	97.27%
KNN (1-Nearest)	0.98	0.99	0.54	96.73%
KNN (3-Nearest)	0.98	0.98	0.54	97.73%

The results in Table II indicate that, the performance of the algorithm increases if user's information is included in training the classification models. Also, SVM showed better accuracy compared to other algorithms. This may be due to the inherently discrete nature of Bangla text morphology and the higher dimensional nature with few irrelevant features. Hence, the accuracy of SVM, which construct a hyper plane or set of hyper planes in a high or infinite dimensional space, on Bangla text justifies the superior performance. Fig. 2 presents comparison of accuracies achieved after applying different classification algorithm.

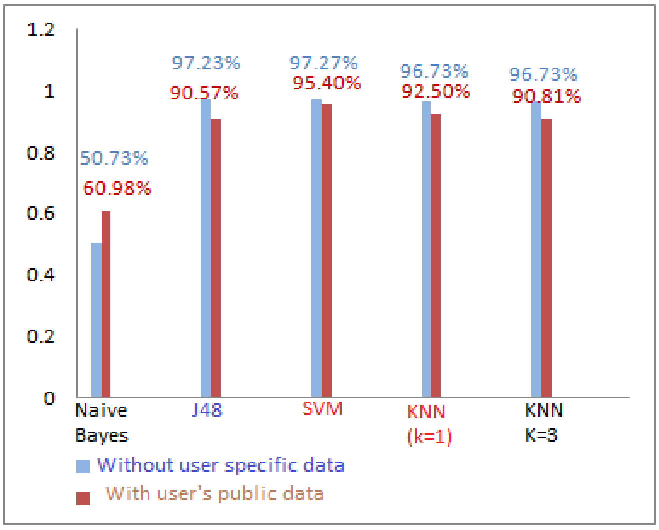

Fig. 2: Accuracy of the proposed cyber bullying detection algorithm achieved after applying different machine learning techniques on Bangla text contents.

TABLE IV
Performance of Social MEDIA BULLYING detection ON ENGLISH using user's posts only

Algorithm	Precision	F1-Score	ROC	Accuracy
Naïve Bayes	0.80	0.39	0.78	40.98%
J48	0.89	0.89	0.92	91.07%
SVM	0.95	0.95	0.94	95.32%
KNN (1-Nearest)	0.87	0.85	0.84	85.10%
KNN (3-Nearest)	0.86	0.86	0.87	87.81%

The results in Table IV shows that, SVM is the best classifier for English text in terms of accuracy. And our research evident that, SVM performs better accuracy for Bangla text also.

V. Conclusion

Though a notable amount of work has been performed for cyber bullying detection on English text, very few work have been done on Bangla text. In this paper revisited four state-of-the-art supervised machine learning algorithms including NB, J48, SVM and KNN on Bangla text and empirically compared their classification performance. An experimental result indicates that the SVM based method achieves best accuracy and the performance improves if user specific data can be included. Due to high-dimensional input space, few irrelevant features and linearly separable nature of text dataset, SVM performs better than other classification algorithm for text classification. In future, significance of individual features can be studied for further enhancement of the method.

References

[1] Rice, Eric, et al. "Cyber bullying perpetration and victimization among middle-school students." *American Journal of Public Health (ajph)*, pp. e66-e72, Washington, 2015.

[2] Bangladesh Telecommunication Regulatory Commission, http://www.btrc.gov.bd/content/internet-subscribers-Bangladesh-january-2018, [Last Accessed on 18 Mar 2018].

[3] Mandal, Ashis Kumar,Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2014.5508.

[4] Dani, Harsh, Jundong Li, and Huan Liu, "Sentiment Informed

Cyberbullying Detection in Social Media" *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2017.

[5] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. "Modeling the detection of Textual Cyberbullying." *The Social Mobile Web* 11.02 (2011): 11-17.

[6] Huang, Qianjia, Vivek Kumar Singh, and Pradeep Kumar Atrey. "Cyber bullying detection using social and textual analysis." *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*. ACM, 2014.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison and L. Edwards, "Detection of Harassment on Web 2.0", *1st Content Analysis in Web 2.0 (CAW 2.0)*, Madrid, Spain, 2009.

[8] Al-garadi, M. A., Varathan, K. D., Ravana, S. D., "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network"*Journal of Computers in Human Behavior,* Vol. 63, pp. 433-443, 2016.

[9] Isa, Sani Muhamad, and Livia Ashianti. "Cyberbullying classification using text mining." *IEEE 1st International conference on Informatics and Computational Sciences (ICICoS),* 2017.

[10] Nitin Rajvanshi, K. R. Chowdhary. "Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA." International Journal of Engineering Research & Technology, *IJERT*, Vol VI, pp. 09, India, 2017.

[11] Wahbeh, Abdullah H., and Mohammed Al-Kabi. "Comparative assessment of the performance of three WEKA text classifiers applied to arabic text." *Abhath Al-Yarmouk: Basic Sci. & Eng* 21.1, pp. 15-28, 2012.

[12] Gogoi, Moromi, and Shikhar Kumar Sarma. "Document classification of Assamese text using Naïve Bayes approach." *International Journal of Computer Trends and Technology* 30.4 (2015): 1-5.

[13] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, pp. 1-47, 2002.

[14] B. Agarwal and N. Mittal, "Text Classification Using Machine Learning Methods A Survey," in Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, 2014, pp. 701-709.

[15] Yin, D., Xue, Z., & Hong, L. (2009). Detection of Harassment on Web 2.0. Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009, 1-7.

[16] K. Reynolds, "Using Machine Learning to Detect Cyberbullying," the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science, pp. 1-4, 2012.

[17] X.-L. Liu, S. Ding, H. Zhu, and L. Zhang, "Appropriateness in applying SVMs to text classification", Comput Eng Sci, vol. 32, pp. 106-108, 2010.

[18] L. Zhijie, L. Xueqiang, L. Kun, and S. Shuicai "Study on SVM Compared with the other Text Classification Methods," in Education Technology and Computer Science (ETCS), 2010 Second International Workshop on, 2010, pp. 219-222.

[19] Nidhi, Vishal Gupta, 2012. "Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach" Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING

[20] R ajan, K., Ramalingam, V.,Ganesan, M., Palanivel, S. and Palaniappan, B "Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural network"In: Expert Systems with Applications, Elsevier, Volume 36 Issue 8, DOI= 10.1016/j.eswa.2009.02.010(2009).