Suspicious Comment Detection from Social Media

Md. Albin Hossain (191-15-12926)

Department of Computer Science and Engineering

Daffodil International University

Dhaka, Bangladesh

albin15-12926@diu.edu.bd

Raju Ahmed (191-15-12672)

Department of Computer Science and Engineering

Daffodil International University

Dhaka, Bangladesh

raju15-12672@diu.edu.bd

**Abstract- With the progressive use of social media the use of suspicious Bangla texts has been increased. Through social media, some people spread hatred and negativity. Significant amount of research has been performed on detecting suspicious text in the English language. Not enough Bangla abusive text detection work has been done. In this study, we applied five classifiers to the dataset to obtain a better result. A large dataset was collected from social media used to detect suspicious comments. After preprocessing and with selected feature extraction, five classifiers MNB, RMFR, SVC, DT and k-NN, and a neural model RNN were applied. In this study, with the full dataset RNN obtained the highest accuracy of 84.93%.**

**Index Terms- Bangla, Suspicious, BLP**

1. **Introduction**

Everyday, a massive volume of Bangla text contents are being produced on social media and other online platforms. Some of them are faked, fabricated, forged, or even suspicious and harmful for society[1]. Suspicious Bangla text contents are increasing day by day with the increase of Bangla users, because of ill-usage of the social media by a few individuals. Some people use social media to share illegal activities, for bullying other people,for smishing, publicizing incitement related contents and so on. Moreover, several extremist users use social media or blogs to spread suspicion and violence which can be considered one kind of threat to national security. Bangla is the 7th most spoken language in the world as around 245 million people speak Bengali as their native language. To ensure the cyberspace security and mitigate national threats the Bengali language suspicious text detection systems are required.

A text could be detected as suspicious if it hurts religious feelings, provoke people against government and law enforcement agencies, motivate people to perform acts of terrorism, perform criminal acts by phishing, instigate a community without any reason, and extortion acts[4]. Detecting suspicious texts manually from the large amount of internet text contents is impossible. That’s why an automatic suspicious text contents detection system should be developed. Responsible agencies will use this tool/system that can automatically detect suspicious text. It will also help to identify threats in cyberspace which will be communicated by Bangla text contents. Automatic detection of suspicious text systems can quickly detect fishy texts. Respective authorities can take immediate measures, which eventually helps to reduce suspicious activities and virtual harassment mediated through online.

NLP can be used to analyze the sentiment of the Bangla text. This allows us to differentiate whether the sentence is suspicious or not[2].

In this study we used the RNN neural model. The neural models primarily represent the variable length text as a fixed length vector. These model layers map n-grams or subword units, words to vector representations to combine different architectures of neural networks. NBOW, RNN, CNN are some models to model text. These models summarize the meaning of embeddings of words in the input text with a fixed length vectorial representation. RNN is popular in NLP problems for it’s very suitable recurrent structure to process the variable length text. A RNN recursively applies a transition function to its internal hidden state vector of the input sequence. The strategy for modeling text sequence is to map the text sequence to a fixed-sized vector using one RNN, and use the vector to a layer for classification.

1. **Literature Review**

For the highly resourced languages like Arabic, English, Chinese and other European languages Suspicious content detection is a well-known research issue. But, only little research activities have been done yet to classify text with suspicious content in the Bangla language processing domain. Not enough meaningful research activities have been conducted yet to classify Bangla text with suspicious content.

Md Faisal et al. [2] used Deep Neural Network for Cyberbullying Detection from Social Media Comments in Bangla Language. They proposed a binary and multiclass classification model using hybrid neural networks for bully expression detection in Bengali language. They collected 44,001 users’ comments of five categories from popular Bangladeshi public Facebook pages. We have examined the performance of our proposed models from different perspectives. They got 87.91% accuracy from binary classification and 85% accuracy with ensemble technique after neural network for multiclass classification.

Omar Sharif created a system to detect suspicious comments or texts in Bengali. The researchers state that they were the setters of the system to investigate Bengali language suspicious text detection. They developed a dataset containing 7000 text documents in Bengali and used multiple algorithms like Logistic regression, Decision tree, stochastic Random forest, gradient descent, and Multinomial Naïve Bayes. Results obtained from these algorithms were compared with each other, existing systems and human experts (baseline). Finally, researchers used those above methods to distinguish a text whether it is suspicious or non-suspicious [1].

A deep learning based tweet classification system was developed using sentiment analysis techniques to classify the tweets as extremist or or non-extremist by Muhammad Zubair Asghar. They proposed a analysis framework for terrorism-related content which does classifying tweets into the classes extremist and non-extremist, based on social media posts on Twitter generated by users. Their experimental results are encouraging and provide a gateway for future researchers[5].

Shahin Akter proposed the use of machine learning algorithms and the inclusion of user information for cyberbullying detection on Bangla text.They collected a set of Bangla text from available social media platforms and labelled the instances as two classes bully and not bully to use for training different ML based models.

Dr.S.Gnanavel created a machine learning model using a logistic regression algorithm to implement the model in a live chat to detect abusive and harmful comments in real time. This study detects types of different toxicity levels and classifies them into severe toxic, insults, obscene, toxic, threat and identity-based hatred and to list out the names of the abusers and toxic users[7].

Md Gulzar Hussain thinks that a little work has been done detecting abusive text in Bangla language. They also think that by using an abusive text detection system it is possible to prevent cyber crimes like blackmailing, harassment on online platforms and cyber bullying. They collected comments from various social sites where people share their sentiment, opinions, views etc.They proposed a root level algorithm to detect abusive text and also proposed unigram string features to get a better result.[8]

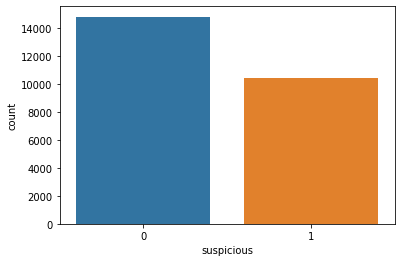
As far as we are aware, none of the remarkable research conveyed so far focuses on detecting suspicious Bangla and shortcut word text. In this study, our goal is to develop a Machine Learning based suspicious Bangla text detection model which is trained on our new dataset with various algorithms.

1. **Workflow**

The paper’s goal is to create a system based on data mining and machine learning that can detect suspicious user comments in real time on social media, micro-blogging sites. It consists of four sub sections such as *A. The Dataset, B. Block Diagram, C.Preprocessing of Text and D. Model Training.*

*A. The Dataset*

The dataset consists of comments collected from different social media popular pages such as popular Bangladeshi Facebook pages comments, popular youtube video comments. After collecting data each of the comments was labeled manually. The dataset contains 25000+ comments and its respective label 0 ro 1. 1 for suspicious and 0 for not suspicious.

**Table 1: Dataset Counts**

*B. Block Diagram*



*C. Preprocessing of Data*

We preprocess the data with 3 steps.

* Remove Punctuations – Punctuation does not contribute to classification. That's why all punctuations will be removed for each text comment.
* Remove stop words – The articles and prepositions are removed using stop word selection. Stopwords like “অতএব”, “অথচ”, “অথবা” etc.
* Stemming– Words will be reduced down to their base form by removing suffixes like “ছে”, “ইয়া”, “য়ে” which are known as stemming.

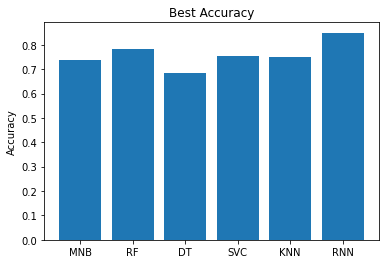
*D. Model Training*

We performed a comprehensive study to find an appropriate model for Bangla text categorization. We considered Multinomial Naive Bayes, Random Forest, SVM, Decision Tree, k-Nearest Neighbor classifier. The dataset was split into an 80:20 ratio for the training set and the test set. All of the classifiers mentioned here were applied in the dataset. Google Colaboratory Environment was used for the working platform and python language was used to implement the programming.

1. **Experimental Result**

| **Algorithms** | **Accuracy** |
| --- | --- |
| Multinomial Naive Bayes | 73.36% |
| Random Forest | 78.82% |
| Decision Tree | 69.40% |
| Support Vector Machines | 74.65% |
| K Neighbors Classifier | 73.60% |
| RNN | 84.93% |

***Comparing the best accuracy:***

****

1. **Conclusion and Future work**

Remarkable amount of work has been done for suspicious text detection on languages such as English, but insufficient work has been done on Bangla. For this research work, a dataset of almost 25,000 instances has been collected from different social sites like YouTube and Facebook. After pre-processing the, several classifiers including MNB, SVM, RMFR classifier, k-NN and decision tree and Neural model RNN were applied. The RNN gave the highest accuracy. There is scope and possibility to improve our approach in this study.

1. **Reference**

[1] Sharif, Omar, Mohammed Moshiul Hoque, A. S. M. Kayes, Raza Nowrozy, and Iqbal H. Sarker. "Detecting suspicious texts using machine learning techniques." Applied Sciences 10, no. 18 (2020): 6527.

[2] Ahmed, Md Faisal, Zalish Mahmud, Zarin Tasnim Biash, Ahmed Ann Noor Ryen, Arman Hossain, and Faisal Bin Ashraf. "Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language." arXiv preprint arXiv:2106.04506 (2021).

[3] Das, Suharto. "Social Media Shortcut Word to Proper English Word." PhD diss., Daffodil International University, 2020.

[4] Bertram, Luke. "Terrorism, the Internet and the Social Media Advantage: Exploring how terrorist organizations exploit aspects of the internet, social media and how these same platforms could be used to counter-violent extremism." Journal for deradicalization 7 (2016): 225-252.

[5] Ahmad, Shakeel, Muhammad Zubair Asghar, Fahad M. Alotaibi, and Irfanullah Awan. "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques." *Human-centric Computing and Information Sciences* 9, no. 1 (2019): 1-23.

[6] Akhter, Shahin. "Social media bullying detection using machine learning on Bangla text." In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 385-388. IEEE, 2018.

[7] Gnanavel, S., N. Duraimurugan, M. Jayalakshmi, M. Rohith, B. Rohith, and S. Sabarish. "A Live Suspicious Comments Detection using TF-IDF and Logistic Regression." Annals of the Romanian Society for Cell Biology (2021): 4578-4586.

[8] Hussain, Md Gulzar, Tamim Al Mahmud, and Waheda Akthar. "An approach to detect abusive Bangla text." In *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, pp. 1-5. IEEE, 2018.

[6] Zam, Dem, Tshewang Dema, and Amol Vasudeva. "Active Chat Monitoring and Suspicious Chat Detection." (2019).

[10] Jahan, Maliha, Istiak Ahamed, Md Rayanuzzaman Bishwas, and Swakkhar Shatabda. "Abusive comments detection in bangla-english code-mixed and transliterated text." In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pp. 1-6. IEEE, 2019.

[11] Mandal, Soumil, Sainik Kumar Mahata, and Dipankar Das. "Preparing bengali-english code-mixed corpus for sentiment analysis of indian languages." *arXiv preprint arXiv:1803.04000* (2018).

[7] Avaghade, S. B., Awale Vinayak Rajkumar, Patil Jaydeep Jayvant, Sonavane Sayali Prithviraj, Patil Aishwarya Dilip, and Karad DACOE. "ACTIVE CHAT MONITORING AND SUSPICIOUS DETECTION OVER INTERNET."

[8] AlGhamdi, Mohammed A., and Murtaza Ali Khan. "Intelligent analysis of Arabic tweets for detection of suspicious messages." Arab J Sci Eng 45, no. 8 (2020): 6021-6032.

[9] Appavu, Subramanian, Muthu Pandian, and R. Rajaram. "Association rule mining for suspicious email detection: a data mining approach." In 2007 IEEE Intelligence and Security Informatics, pp. 316-323. IEEE, 2007.

[10] Ahmed, Md Faisal, Zalish Mahmud, Zarin Tasnim Biash, Ahmed Ann Noor Ryen, Arman Hossain, and Faisal Bin Ashraf. "Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language." arXiv preprint arXiv:2106.04506 (2021).

[11] Ahmed, Hanaa Mohsin, and Noor Basim Bethoon. "Cybercrime: Suspicious Viber Messages Detection Model." Int. J. Sci. Eng. Res 8 (2017): 1496-1502.

[12] alias Balamurugan, S. Appavu, Dr R. Rajaram, G. Athiappan, and M. Muthupandian. "Data mining techniques for suspicious email detection: A comparative study." In IADIS European Conference Data Mining. 2007.

[13] Rahman, Md, and Emon Kumar Dey. "Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation." Data 3, no. 2 (2018): 15.

[14] Mujtaba, Ghulam, Liyana Shuib, Ram Gopal Raj, and Roshan Gunalan. "Detection of Suspicious Terrorist Emails Using Text Classification: A Review." Malaysian Journal of Computer Science 31, no. 4 (2018): 271-299.

[15] Gupta, Madhav, Romit Shriwal, Divyansh Negi, Devarshi Pandey, and Ms Shikha Rai. "Suspicious Mail Detection." (2019).

[16] Nizamani, Sarwat, Nasrullah Memon, Uffe Kock Wiil, and Panagiotis Karampelas. "Modeling suspicious email detection using enhanced feature selection." arXiv preprint arXiv:1312.1971 (2013).

[17] Al Mansoori, Saeed, Afrah Almansoori, Mohammed Alshamsi, Said A. Salloum, and Khaled Shaalan. "Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis." TEM Journal 9, no. 4 (2020): 1313.

[18] Rajaram, Ramasamy, and Appavu Balamurugan. "Suspicious Email detection via decision tree: A data mining approach." Journal of computing and information technology 15, no. 2 (2007): 161-169.

[19] Tagami, Tsubasa, Hiroki Ouchi, Hiroki Asano, Kazuaki Hanawa, Kaori Uchiyama, Kaito Suzuki, Kentaro Inui et al. "Suspicious news detection using micro blog text." arXiv preprint arXiv:1810.11663 (2018).

[20] Chourasia, Anant, Mohit Agrawal, Sakshi Thakur, Ayushi Bais, Muskan Singh, and Vrushali Bongirwar. "Suspicious Online Forum Data Detection." Helix 10, no. 04 (2020): 457-461.