Implementation of Machine Learning to Detect Hate Speech in Bangla Language

Shovon Ahammed[1], Mostafizur Rahman[2], Mahedi Hasan Niloy[3] and S.M. Mazharul Hoque Chowdhury[4]
[1,2,3,4]Department of CSE, Daffodil International University, Dhaka, Bangladesh
E-mail: [1]shovon15-7671@diu.edu.bd, [2]mostafizur15-7764@diu.edu.bd, [3]mahedi15-7763@diu.edu.bd
[4]mazharul2213@diu.edu.bd

Abstract— **Hate speech is a crime in all countries. Hate speech can be for women, religions, countries, cultures. The big problem for hate speech is that it entices the evil people. Moreover, it inspires them to spread hatred in the society. Bangla is one of the topmost spoken languages in the world. But hate speech detection in Bangla language is rare. Our purpose is to detect hate speech in Bangla language. To perform the task, we were in need of the Bangla datasets. But the Bangla dataset is not available. So, we have collected data from Facebook. Collecting data from the social site is very hectic. The data contain mixed languages, grammatical mistakes. So, we made a team to collect the data. Another team was to process the data. And finally, we labeled the data as hate speech or not. The team members had enough knowledge about hate speech. They were neutral towards the data. Our data contain hate speech against women, community, culture, ethnicity, race, sex, disability. Machine Learning approach is ideal for our work. We have used the SVM and Naïve Bayes algorithm for our work and got a maximum accuracy of 72%.**

Keywords— SVM, Machine Learning, Supervised Learning, Naïve Bayes, Hate Speech.

I. Introduction

Bangla is the state language of Bangladesh along that millions of people speak in Bangla as their first language. Bangla has come from Sanskrit. Our language has age-old tradition and culture. We are the only country that gave blood for language. Our language has a value, and we have to respect that. We should not use this language for wrong purpose. But it is a matter of great sorrow that there are many cases of hate speech in the Bangla language.

Hate speech is a global problem. People of different parts of the world are now connected to each other with the help of the internet. Now people can share their feeling with the whole world in the blink of an eye. At the same time, people can share hatred along with speech in a second which is a matter of concern. We know hatred hurts people mentally and it bears a deep impact in their life. The more people are getting the opportunity to share their thoughts and views on the different matter the more incident of hate speech of taking place. The more the hate speech spreads the more damage it does. These hate speeches are a threat to our unity. Hate speeches are dividing us.

We have made the dataset of hate speech along with regular speech. We have collected data from Facebook. It is one of the most popular social networking sites in Bangladesh. Millions of people in Bangladesh use Facebook. And most of them use Bangla language on Facebook. The Internet pack is getting cheaper day by day. The number of internet users is increasing. Now people of every part of Bangladesh are using Facebook. They are giving their analysis and views on different sectors. In this procedure there is a clash of thinking, liking, disliking. Whenever there is a clash, they are using tons of hate speech against each other. So, they use Bangla Language to create a post or to comment on a post. People are attacking each other for different reasons. People have become so intolerant. For this reason, we have chosen Facebook to collect data.

In this work, we built a new dataset to find malice in Bangla Language that contains hate speech on different categories such as religion, community, gender, race.

We have used web scrapper to scrap the data. We have labeled our data in two categories either it is a hate speech or not. The contribution of our work is to creating a new dataset for hate speech detection in Bangla Language and applying the algorithm to detect it.

II. Related Work

Natural Language Processing is playing huge role in detecting hate speech. It is very efficient to detect hate speech with NLP.

Axel Rodríguez et al. [1] have used four steps to detect hate speech in the English language. They have used data from Facebook. Their first step is the discovery stage. In this step, they had identified some pages which produce hate speech. Not all the post of the pages contains hate speech, so they had used a filter to sperate hate speech from regular speech. They have also done sentimental analysis. They have used Valence Aware Dictionary for sentiment reasoning.

Latent Semantic Analysis is a very popular natural language processing method. Ilham Maulana Ahmad et al. [2] have used the LSA method based on the image. Their approach was to extract information from the image. They

Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

have used data mining. To gather information, they have used twitter. They have got an average accuracy of 57.9%.

Ricardo Martins et al. [3] have used emotional words to classify hate speech. They have obtained an accuracy of 80.56% with a support vector machine.

The fast Text approach is a great method to find malice sentences. In this paper, Nur Indah Pratiwi et al. [4] have used Fast text approach to detect hate speech in Instagram comments. They have used word n-gram and char n-grams.

Arum Sucia Saksesi et al. [5] used a recurrent neural networks to detect hate speech. They have used the Twitter API to collect the data. For text analysis they have performed case folding, tokenizing, cleaning, stemming. They made a combination of LSTM with RNN. To find the result of the LSTM hidden layer they have used softmax regularization. They partitioned the data with different ratios at different times. At learning rate 0.007 they got an accuracy of 93.8%, precision of 92% and recall of 93%.

N.D.Gitari et al. [6] have used lexicon in their work. They got average results with the lexicon model. Their F-score was 70.83.

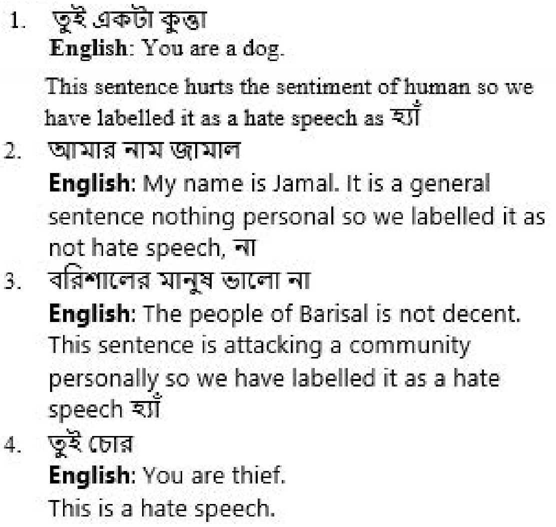
Erryan Sazany et al. [7] have detected hate speech using deep learning. Trisna Febriana and Arif Budiarto [8] have detected hate speech in the Indonesian language.

Researchers have applied different methods to classify hate speech. Most of them have used social sites to collect data. They have used a machine learning-based approach.

III. Methodology

This paper aims to find hate speech in Bangla language. We have taken a machine learning approach for the task. Now we are going to describe our work. The main challenge of our work was to build the dataset.

Example:



1. তুই একটা কুত্তা
   English: You are a dog.
   This sentence hurts the sentiment of human so we have labelled it as a hate speech as হ্যাঁ
2. আমার নাম জামাল
   English: My name is Jamal. It is a general sentence nothing personal so we labelled it as not hate speech, না
3. বরিশালের মানুষ ভালো না
   English: The people of Barisal is not decent. This sentence is attacking a community personally so we have labelled it as a hate speech হ্যাঁ
4. তুই চোর
   English: You are thief.
   This is a hate speech.

A. Forming the Dataset

Data acquisition and data annotation are the two main parts of our data formation.

1). Data Acquisition

Facebook is one of the biggest social network platforms. It generates tons of data every day. People of all ages and groups use Facebook. So, we decided to take data from Facebook. Scraper helps us to get data from websites. We have used web scraper to get the data from Facebook. As we have used web scraper the scraper scraped different types of data. The dataset had good as well as bad comments. Some comments were targeting women, religious groups. While some comments were targeting race, community. And there were normal comments like people were giving their opinion on various subjects with a positive mind.

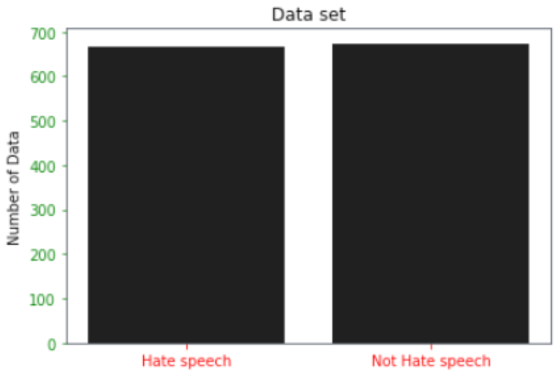
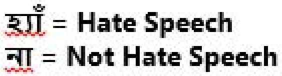


Fig. :1 Dataset Visualization

2). Data Annotation

Data annotation was the key part of our work. We were focused to annotate our data correctly. As our work finds out a sentence is hated speech or not so we made two categories. One category was hating speech and another category was general speech. Then we labeled the data with tags. If the comment has something inappropriate, we labelled that as হ্যাঁ and if the comment is appropriate, we labelled that as না.



হ্যাঁ = **Hate Speech**
না = **Not Hate Speech**

We worked in groups to annotate the data. To be fair with the data the annotation was done in three steps. At the first step, the data was labelled with one group. Then the authenticity of the label was checked by another group. Finally, the ultimate labelling was done with the collaboration of two groups. We worked that way so that we do not commit any mistake on labelling data.

For each comment the answer whether the comment is hated speech or not.

B. Hate Speech Identification

Now to identify the hate speech we have worked in four steps

- Pre-processing
- Data Analysis
- Feature Extraction

Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

- Implementation of Machine Learning

1). *Pre-processing*

Pre-processing means processing the data according to need. We have collected data from Facebook. Without pre-processing the data will not be able to perform well. And for our work data pre-processing, is vital. So, the data had different types of issue. In Facebook comment, people use different types of emoji. In our machine learning-based classification we cannot work with emoji.

So manually we have removed emoji from different comments. Then people perform different types of spelling mistakes while commenting on Facebook.

We tried to correct collect the spelling. While working with data negation handling is an important task. So, we have worked with negation. Performing these operations makes our data prepare for the next steps. After completing these steps, the pre-processing of our data completes.

2). *Data Analysis*

Data analysis is used to get knowledge about the data. Every data is significant. They have their own patterns and value. For example, spam data and ham data have variations with text length. But while we have analyzed our dataset, we have found that text length does not have any significance to categorize the data.

At figure 2 we can see a histogram for text length of hate speech. Text length of 25-40 contains most hate speech. The hate speech length varies but at that points the number of hate speeches is maximum. There is a decent number of hate speeches between the text length of forty-one and ninety-five. The amount of hate speech at the text length of hundred is surprisingly low. The maximum text length of a hate speech is two-hundred and sixty. But one thing is clear most of the hate speech length is around ten to a hundred.

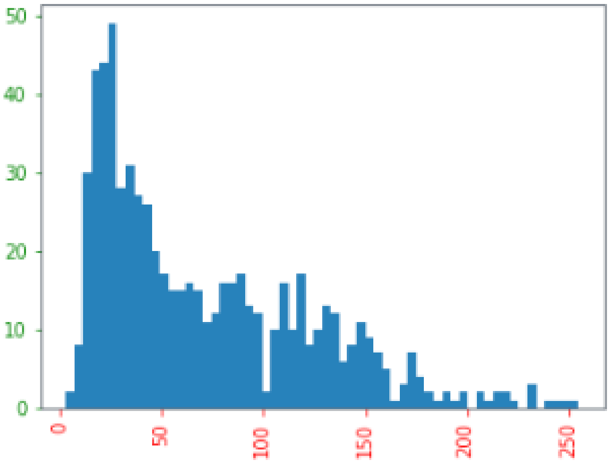


Fig. 2: Text length of Hate Speech

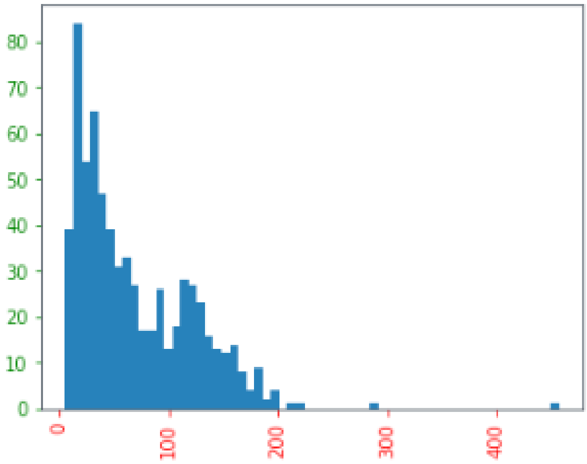


Fig. 3: Text Length of not Hate Speech

At figure 3 is the histogram of neutral speech. The maximum number of neutral speeches is at the text length of ten. And this is very common because there are some common and popular sentences like nice pictures, beautiful, good morning, all the best, happy birthday. Though we, have given the example in English because the text length is almost the same in Bangla. After comparing the histogram of hate speech data and not hate speech data, we found that the text length of not neutral speech data is short. Most of the text length of neutral speech is between one and two hundred.

Table 1: Dataset Distribution

Number of Sentence	Number of Hate Speech	Number of Normal Speeches
1339	665	674

We collected more than 5,000 data to perform the task. Most of them were neutral comments. For this reason, to keep a balance between the data, we made a dataset of 1339 data. At table 1 we can see that in our dataset there are 665 hate speeches and 674 neutral speech. In those hate speeches, people were sharing hatred. On the contrary, in neutral speech people were congratulation each other on different occasions. They were giving suggestion to each other on different topics.

3). *Feature Extraction*

We have extracted the feature with count vectorizer and Term frequency-inverse document frequency vectorizer. The count vectorizer tokenizes the text and creates a vocabulary of known words. Count vectorizer encodes a new document using that vocabulary.

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \tag{1}$$

Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

Here,

$tf_{i,j}$ = number of occurrences of i in j

dfi = number of documents containing I

N = total number of documents

The Term frequency-inverse document frequency is the number of times a word appears in a document divided by the total number of words in that document and the second term is Inverse document frequency, computed as the logarithmic of the number of the documents in the corpus divided by the number of documents where the specific term appears.

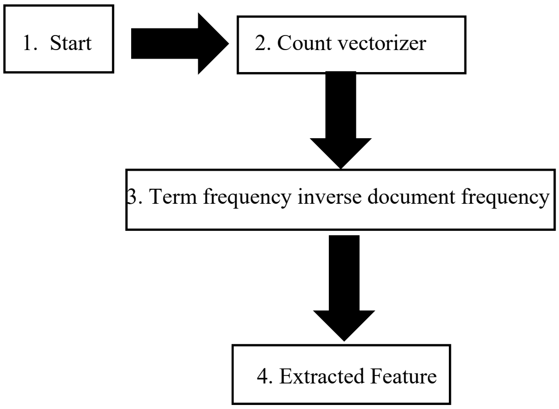


Fig. 4: Flow Diagram of Feature Extraction

4). Implementation of Machine Learning

We have implemented Machine Learning to perform our task. Supervised learning has been used. Supervised learning means supervising the data. As we have collected data from Facebook and then we have labelled the data according to their criteria. After that, we have trained our model with the labelled data. Our model has learnt from the labelled data which are hate speech, and which is not. For classification, we have used two algorithms Naïve Bayes and Support Vector Machine. We divided our data into a training set and testing set. Then we fed the data to both algorithms. After that, we got accuracy, precision, recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

Precision is the value of true positive divided by the summation of the true positive and false positive. The recall is true positive divided by the summation of the true positive and false positive. Accuracy is the summation of the true positive and true negative divided by the summation of true positive, true negative, false positive and false negative

IV. Results

In machine learning, the performance is shown in the form of confusion matrix to F-measures. Recall is true positive divided by the summation of true positive and false negative. And accuracy is summation of a true positive and true negative divided by summation of true positive, true negative, false positive and false negative. After applying SVM we got accuracy of 70% and

Table 2: Test Results with SVM

	Precision	Recall	F1-score
না	0.73	0.70	0.71
হ্যাঁ	0.68	0.70	0.69

After applying Naïve Bayes, we got accuracy of 72% and

Table 3: Test Results with Naïve Bayes

	Precision	Recall	F1-score
না	0.75	0.71	0.73
হ্যাঁ	0.70	0.74	0.72

V. Conclusion

In our work, we made a new dataset in the Bangla language. We divided the dataset into two groups and labeled them. There were anomalies in the dataset, and we processed them to remove the anomalies. After that, we have extracted features from our dataset to use in our model. Then we applied the Support vector machine and Naïve Bayes machine learning algorithms. Both the algorithm performed well with our dataset. We showed Precision, Recall and F1-score for both of the algorithms. Naïve Bayes gave us an accuracy of 72%.

References

[1] Axel Rodríguez, Carlos Argueta and Yi-Ling Chen*, "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis," International Conference on Artificial Intelligence in Information and Communication, pp. 169 – 174, 2019.

[2] Ilham Maulana Ahmad Niam, Budhi Irawan, Casi Setianingsih and Bagas Prakoso Putra, "Hate Speech Detection Using Latent Semantic Analysis (LSA) Method Based on Image," International Conference on Control, Electronics, Renewable Energy and Communications, pp. 166–171, 2018.

[3] Ricardo Martins, Marco Gomes, Jos´e Jo˜ao Almeida, Paulo Novais and Pedro Henriques, "Hate speech classification in social media using emotional analysis," 7th Brazilian Conference on Intelligent Systems, pp. 61–66, 2018.

[4] Nur Indah Pratiwi, Indra Budi, and Ika Alfina, "Hate Speech Detection on Indonesian Instagram Comments using FastText Approach," International Conference on Advanced Computer Science and Information Systems, pp. 447–450, 2018.

[5] Arum Sucia Saksesi, Muhammad Nasrun and Casi Setianingsih, "Analysis Text of Hate Speech Detection Using Recurrent Neural Network," International Conference on Control, Electronics, Renewable Energy and Communications, pp. 242-248, 2018.

[6] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A Lexicon-based Approach for Hate Speech Detection," *Int. J. Multimed. Ubiquitous Eng.*, vol. 10, no. 4, pp. 215–230, 2015.

[7] Erryan Sazany and Indra Budi, "Deep Learning-Based Implementation of Hate Speech Identification on Texts in Indonesian: Preliminary Study," International Conference on Applied Information Technology and Innovation, pp. 114-117, 2018.

[8] Trisna Febriana and Arif Budiarto, "Twitter Dataset for Hate Speech and Cyberbullying Detection in Indonesian Language," International Conference on Information Management and Technology, pp. 379-382, 2019.