IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AN IN-DEPTH SURVEY OF HATE SPEECH DETECTION ON TWITTER USING MACHINE LEARNING

[1]Sneha Gujar, [2]Ankleshwar Vishwakarma, [3]Niharika Solanki, [4]Pratik Upadhyay, [5]Nikhil Pandhare

[1]Department of Computer Engineering,

Genba Sopanrao Moze College of Engineering Balewadi, Pune-411045, Maharashtra, India

Abstract: People from various psychological and cultural backgrounds began communicating more directly with one another as social networks and microblogging websites grew quickly. This led to an increase in "cyber" conflicts between these individuals. As a result, hate speech is used more and more, to the point where it is starting to seriously affect these public areas. The term "hate speech" describes the use of harsh, violent, or offensive language directed towards a particular group of people who have something in common, such as their gender (sexism), race or ethnicity (racism), belief and religion, etc. Although the majority of online social networks and microblogging platforms prohibit the use of hate speech, dur to their massive scale, it is nearly impossible to master all of their material. As a result, it becomes necessary to automatically gathered from the foundation of the methodology. A machine learning algorithm is later trained using these patterns and unigrams among other features. Our tests on a test set of 2,010- tweets demonstrate that our method achieves an accuracy of 87.4 classification and 78.4 classification respectively.

Index Terms - Hate speech recognition, machine learning, Classification, NLP, Sentiment Analysis, Deep Learning.

I. INTRODUCTION:

Bangladesh's official language is Bangla, which is also the mother tongue of millions of people. Sanskrit is where Bangla originated. Our language is steeped in history and culture. Only our nation donated blood in exchange for language. We must respect the value that our language holds. Use of this language improperly is not appropriate. However, the fact that hate speech occurs frequently in Bangla is a cause for deep regret.

Hate speech is a widespread issue. Thanks to the internet, people all over the world can now communicate with one another. Now a days, a person only needs to blink to share their emotions with the entire world. In addition, it is concerning that people can spread hate speech and hatred in the same breath. We are aware that hatred has a profound effect on people's lives and damages people mentally. The number of incidents of hate speech increases as more people have the chance to express their opinions on various topics. The harm caused by hate speech increases with its dissemination. Our unity is in danger because of these hate speech acts. We are divided by hate speech.

Along with regular speech, we have also created a dataset of hate speech. We have Facebook data that we have gathered. In Bangladesh, it is among the most widely used social networking sites. Facebook is used by millions of people in Bangladesh. Additionally, the majority of them speak Bangla on Facebook. Every day, the Internet pack becomes more affordable. The amount of people using the internet is rising. People use Facebook nowadays all over Bangladesh. They are sharing their opinions and analyses of various industries. There is a conflict between liking and disliking during this process. They constantly use hate speech against one another during conflicts. Thus, when they write a post or leave a comment, they utilize Bangla language.

Humans are disparaging one another for various motives. People are now so incredibly intolerable. We have decided to gather data on Facebook as a result.

We created a new dataset in this work to identify malice in Language in Bangla that uses hate speech on various categories like race, gender, community, and religion. To scrape the data, we used a web scraper.

As we've classified our data into two groups: either it contains hate speech either way. Our work has contributed to the development of a new dataset for the identification of hate speech in Bangla putting the algorithm to work to find it.

II. LITERATURE SURVEY:

1. **Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions an Perform Hate Speech Detection "Hajime Watanabe, Mondher Bouazizi, Tomoaki Ohtsuk"**

People from various cultural and psychological backgrounds began to communicate more directly with one another as social networks and microblogging websites grew quickly. This led to an increase in the number of cyber conflicts between these individuals. As a result, hate speech is used more and more, to the point where it is starting to seriously affect these public areas. The term "hate speech" describes the use of aggressive, violent, or offensive language directed towards a particular group of people who have something in common, such as their gender (sexism), race or ethnicity (racism), beliefs and religion, etc. Although hate speech is prohibited on the majority of online social networks and microblogging websites, it is nearly impossible due to the scale of these platforms. to be in charge of all their content. As a result, it becomes necessary to automatically identify this type of speech and filter any content that uses offensive or inciting language. In this paper, we suggest a method for identifying hate speech on Twitter. Our methodology relies on automatically gathered unigrams and patterns from the training set. A machine learning algorithm is later trained using these patterns and unigrams among other features. Our tests on a test set of 2,010-tweets demonstrate that our method achieves an accuracy of 87.4 classification and 78.4 accuracy.

2. **Implementation of Machine Learning to Detect Hate Speech in Bangla Lan- guage Shovon Ahammed1, Mostafizur Rahman2, Mahedi Hasan Niloy3 and S.M. Mazharul Hoque Chowdhury4 1,2,3,4Department of CSE, Daffodil Interna- tional 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**

It is illegal to spew hate speech in any country. Hate speech can be directed towards women, countries, religions, or cultures. Hate speech is a major problem because it attracts evil people. Additionally, it motivates them to incite hatred within the community. One of the most widely spoken languages in the world is Bengali. Nonetheless, it is uncommon to find hate speech in Bangla. Our goal is to identify hate speech in Bengali. We needed the Bangla datasets in order to complete the task. However, the Bangla dataset isn't accessible. Thus, we have gathered information from Facebook. It's very busy gathering data from the social media platform. There are grammatical errors and mixed languages in the data. We thus formed a team to gather the information. Another group aimed to handle the data. Lastly, we classified the information as hate speech or not. The group members were sufficiently knowledgeable about hates speech. They had no opinion about the data. Hate speech against women, the community, culture, ethnicity, race, sex, and disability is present in our data. The machine learning method is perfect for the work we do. Using SVM and the Naive Bayes algorithm, we achieved a maximum accuracy of 72 keywords. – Hate Speech, Naive Bayes, SVM, Machine Learning, Supervised Learning.

3. **Parasitic Gate Resistance Impact on Triple-Gate FinFET CMOS Inverter Edgar Solis Avila, Julio C. Tinoco, Senior Member, IEEE, Andrea G. Martinez- Lopez, Member, IEEE, Mario Alfredo Reyes-Barranca, Antonio Cerdeira, Se- nior Member, IEEE, and Jean-Pierre Raskin, Fellow, IEE**

In this work, we examine the effect of Fin FET gate resistance on the inverter and ring oscillator performance using a complete intrinsic–extrinsic model for symmetric doped double-gate MOSFET. It is demonstrated that the multi finger configuration, which results in a decrease in gate resistance, can improve the propagation delay when the total number of fins stays constant. Moreover, reducing source/drain fin extension length and fin spacing are crucial for enhancing the performance of digital circuits. Index Terms: Fin FETs, digital circuits, high-speed performance, CMOS ring oscillator (RO), CMOS inverter, and par-asities.

III. METHODOLOGY:

The purpose of this essay is to identify hate speech in Bangla. For this task, we have adopted a machine learning methodology. We will now give a description of our work. The creation of the dataset presented our work's biggest obstacle.

Example:

1. তুমি একটা কুকুর
   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 is as not hate speech, না.
3. বরিশালের মানুষ ভদ্র নয়
   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. তুমি চোর
   English: You are thief.
   This is a hate speech হ্যাঁ.

A. *Forming the Dataset-*

The two primary components of our data formation are data annotation and data acquisition.

1) *Data Acquisition-*

Among the most popular social networking sites is Facebook. Every day, tons of data are generated by it. Every age group and demographic use Facebook. Thus, we made the decision to use Facebook's data. We can obtain data from websites with the aid of scrapers. Web scrapers were used by us to obtain the Facebook data. The web
scraper that we used collected various kinds of data. The dataset included both positive and negative remarks. Certain remarks were directed at religious and/or female groups. While some remarks attacked communities based on race. Additionally, there were typical remarks made by people sharing their positive viewpoints on a range of topics.

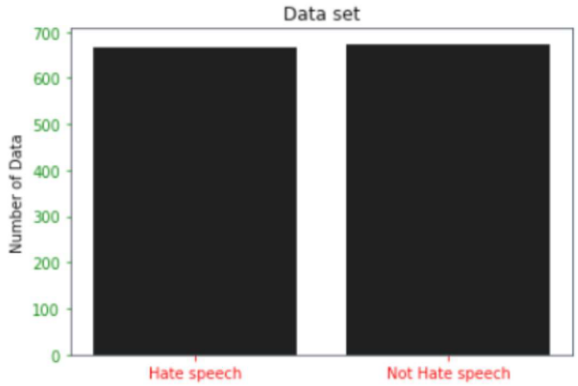

Fig.: Dataset Visualization

2) Data Annotation-

Annotating data was a crucial aspect of our work. Our goal was to accurately annotate our data. In order to distinguish between hate speech and non-hate speech, our work creates two categories. Hate speech fell into one category, and general speech fell into another. After that, we attached tags to the data. If there is anything improper in the comment, it is marked as হ্যাঁ and if the comment is appropriate, we labelled that as না.

(হ্যাঁ = Hate Speech, না = Not Hate Speech)

Our groups' task was to annotate the data. Three steps were taken in order to fairly annotate the data. The data was first labelled with a single group. Then, a different team verified the label's authenticity. Finally, two groups worked together to complete the final labelling. We

conducted our work in this manner to ensure accurate data labelling. The response to each comment, indicating whether or not it constitutes hate speech.

B. *Hate Speech Identification-*

Now, let's take four steps to identify the hate speech.

- Pre-processing
- Data Analysis
- Feature Extraction
- Implementation of Machine Learning

1) *Pre-processing-*

Pre-processing is the process of handling data based on requirements. We have Facebook data that we have gathered. The performance of the data will be compromised without pre-processing. Pre-processing data is also essential for our work. Thus, the data contained a variety of problems. People use a variety of emoji in Facebook comments. Emojis are not compatible with our machine learning-based classification system.

Emojis have therefore been manually eliminated from various comments. Subsequently, individuals commit various spelling errors when leaving comments on Facebook. We tried to fix the spelling in collect. Negation handling is crucial when working with data.

Consequently, we have used negation. By carrying out these procedures, our data becomes ready for the following stages. These procedures bring our data's pre-processing to a close.

2) *Data Analysis-*

The purpose of data analysis is to learn more about the data. Every piece of data matter. They are valued according to their own patterns. For instance, the length of the text affects spam and ham data differently. However, after analysing our dataset, we discovered that there is no meaningful relationship between text length and data classification.

Figure 2 displays a hate speech text length histogram. Most hate speech is found in texts that are 25–40 words long. The duration of hate speeches varies, but at that point, there are no more hate speeches than there can be. Between 41 and 95 text lengths, there are a respectable amount of hate speeches. Surprisingly little hate speech is found in texts longer than 100 words. The highest a hate speech has two hundred and sixty words. However, it is evident that the majority of hate speech is between ten and one hundred words long.

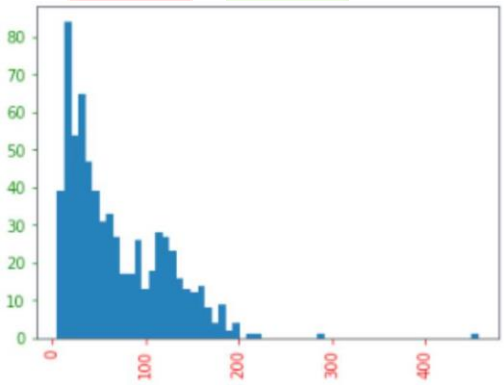

Fig.: Text length of Hate Speech

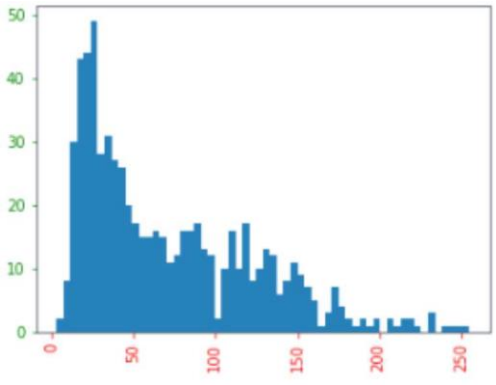

Fig.: Text Length of not Hate Speech

At fig 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.

3) Feature Extraction-

Using the count vectorizer and the term frequency-inverse document frequency vectorizer, we were able to extract the feature. The text is tokenized by the count vectorizer, which also generates a list of commonly used terms. Count vectorizer uses that vocabulary to encode a new document.

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

Here,

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

df_i = Number of documents containing I

N = total number of documents

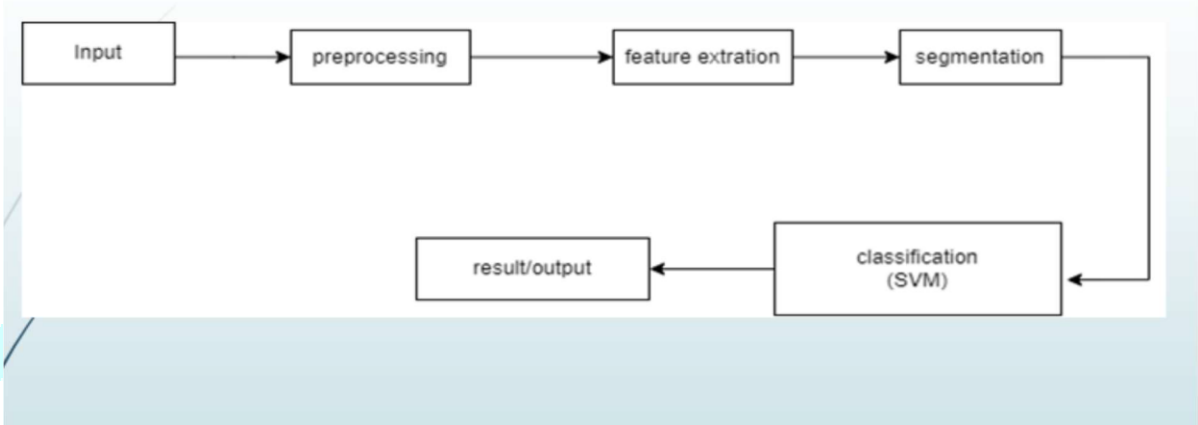

Fig.: Data Flow diagram 2 level

4) Implementation of Machine Learning-

To complete our task, we have put machine learning into practice. There has been use of supervised learning. Monitoring the data is the essence of supervised learning. As we have gathered information from Facebook and labelled it based on their standards. Following that, we used the labelled data to train our model. Through labelled data, our model has been trained to distinguish between hate speech and non-hate speech. We have employed two algorithms for classification: Support Vector Machine and Naïve Bayes. Our data was split into training and testing sets. Next, we fed the information to each algorithm. We then obtained recall, accuracy, and precision.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{..... (2)}$$

$$\text{Recall} = \frac{TN}{TN+FN} \quad \text{..... (3)}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (4)}$$

The calculation of precision involves dividing the true positive value by the total of the true positive and false positive. True positive divided by the total of true positives and false positives is the recall. The calculation of accuracy involves dividing the total of true positive and true negative by the total of true positive, true negative, false positive, and false negative.

IV. SYSTEM DESIGN:

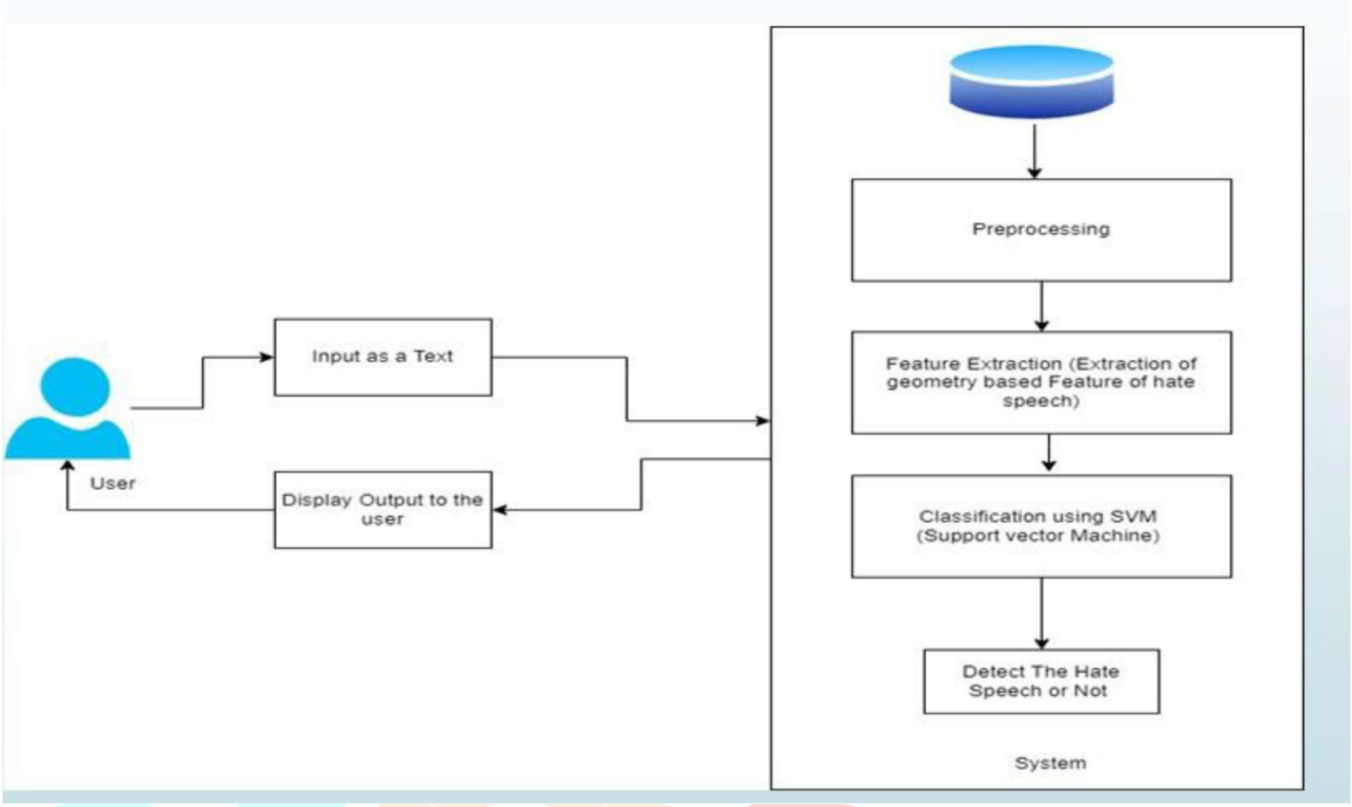

Fig.: System Architecture

V. 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 Fast Text 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