IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, March 2024

Hate Speech and Abusive Language and Abusive Language Detection in Twitter using Machine Learning

Sakshi Dhatrak[1], Sanika Rasal[2], Sakshi Bodke[3], Sakshi Shinde[4], Prof. Miss. Aishwarya Sanap[5]

Students, Department of Information Technology[1,2,3,4]
Professor, Department of Information Technology[5]
Matoshri Aasarabai Polytechnic, Eklahare, Nashik, Maharashtra, India

Abstract*:* *Twitter's central goal is to enable everybody to make and share thoughts and data, and to communicate their suppositions and convictions without boundaries. Twitter's job is to serve the public discussion, which requires portrayal of a different scope of points of view. Yet, it does not advance viciousness against or straightforwardly assault or undermine others based on race, nationality, public cause, rank, sexual direction, age, inability, or genuine illness. Hate speech and abusive language can hurt a person or a community. So, it is not appropriate to use hate speech and abusive language. Now, due to increase in social media usage, hate speech and abusive language is very commonly used on these platforms. So, it is not possible to identify hate speech and abusive languagees manually. So, it is essential to develop an automated hate speech and abusive language detection model and this research work shows different approaches of Natural Language Processing for classification of Hate speech and abusive language through Machine Learning Algorithms.*

Keywords: Logistic Regression, SVM, Tf-Idf, Random Forest, Hate speech and abusive language

I. INTRODUCTION

Due to increasing scale of social media, people are using social media platforms to post their views. Giving opinions which are harsh or rude to someone directly on face is a difficult task. So, people feel it is safe over internet to abuse or post something offensive to others. So, they feel secured posting such content on the internet. Due to this the use of hate speech and abusive language over the social media is increasing daily. So, as to handle such a large data of users over social media, automatic detection of hate speech and abusive language methods are required. In this paper we use machine learning methods to classify whether hate speech and abusive language or not. There are a number of machine learning applications, One of them is for text based classification. Each instance or here we can say each tweet is represented using the same set of features used by machine learning algorithms. There are two types of problems solved machine learning algorithms, supervised and unsupervised. Supervised learning is the task of training model based on given dataset containing both set of features and labels. Though unsupervised learning is the training system function in which data set is neither categorized nor named. Supervised learning is further divided into two types regression and classification based on labels of dataset. Here we concerned only about classification. Classification machine algorithms used categorical dataset and are used to classify the class/category of the unknown instance

II. LITERATURE SURVEY

Hate speech and abusive language on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate speech and abusive language Detection" Hajime Watanabe, Mondher Bouazizi, Tomoaki Ohtsuk"

With the rapid growth of social networks and microblogging websites, communication between people from different cultural and psychological backgrounds became more direct, resulting in more and more "cyber" conflicts between these people. Consequently, hate speech and abusive language is used more and more, to the point where it became a serious problem invading these open spaces. Hate speech and abusive language refers to the use of aggressive, violent

DOI: 10.48175/IJARSCT-15677

or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or their believes and religion, etc.

Implementation of Machine Learning to Detect Hate speech and abusive language Moreover, it inspires them to spread hatred in society. Bangla is one of the topmost spoken languages in the world. But hate speech and abusive language detection in Bangla language is rare. Our Parihar et al., 2021 [18] Hate speech detection is a very difficult task and continues to be a societal problem. There is a very fine line between what is a hate speech and what is not. For example, a satire might also be considered as a possible threat but it is not actually a hate speech. The annotation and collection of data for building a model for hate speech detection is thus a very troublesome task. As discussed, this problem can be solved by narrowing down the criteria for annotations. Similarly, there is a need to focus research on code-mixed languages and regional languages as well. Language models and deep learning models have shown promising results in hate speech classifications. For tackling with unbalanced data, the up sampling or down sampling techniques based on language models should be researched upon. The challenges discussed above must be tackled with more research in the domain so that the internet becomes more inclusive, welcoming and free from hate.

Mahibha et al, [13] They concluded that, from the output obtained from the different models it could be inferred that deep learning models outperform the machine learning models considering the offensive language classification problem for the data set provided by HASOC@FIRE-2021 for Task 1 associated with code mixed Tamil. Among the deep learning model transformer-based models has done the more accurate predictions compared to recurrent models, hence more scope for transformer-based models could be identified for research based on Dravidian languages and in specific Hate and Offensive language-based researches.

Zeerak Waseem et al. [20] classify the hate speech on twitter. In their research, they employed character Ngrams feature engineering techniques to generate the numeric vectors. The authors fed the generated numeric vector to the LR classifier and obtained overall 73% F-score. While, Chikashi Nobata et al. [6] used the ML -based approach to detect the abusive language in online user content. In their research authors employed character Ngrams feature representation technique to represent the features. The authors fed the features to the SVM classifier. The results showed that the classifier obtained overall 77% F-score. Shervin Malmasi et al [14] used an ML -based approach to classify hate speech in social media. In their research, the authors employed 4grams with character grams feature engineering techniques to generate numeric features. The authors fed the generated numeric features to the SVM classifier. The authors reported maximum of 78% accuracy.

purpose is to detect hate speech and abusive language 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.

Dynamic weighted attention with multi channel convolutional nural network for Emotion Recognition.In this paper, based on a full intrinsic–extrinsic model for symmetric doped double-gate MOSFET, we analyze the impact of FinFET gate resistance over the inverter and ring oscillator performance. It is shown that, when the total number of fins remains constant, the propagation delay can be improved thanks to the multifinger configuration that translates into the gate resistance reduction.

Syntactic Edge -Enhanced Graph Convlutional Networks for Aspech level sentiment classification with interactive attention.In natural language processing, aspect-level sentiment classification is a popular research area (NLP). How to create efficient algorithms to model the relationships between aspects and opinion words that appear in a sentence is one of the major challenges. The graph convolutional networks (GCNs) achieve the promising results among the various methods suggested in the literature because of their strong ability to capture the large distance between the aspects and the opinion words. Sentence-Level Classification using Parallel fuzzy deep learning classfiier:In this paper the author specifies the classification using parallel fuzzy deep learning classifier. At present, with the growing number of Web 2.0 platforms such as Instagram, Facebook, and Twitter, users honestly communicate their opinions and ideas about events, services, and products. Owing to this rise in the number of social platforms and their extensive use by people, enormous amounts produced hourly.

III. PROPOSED SYSTEM

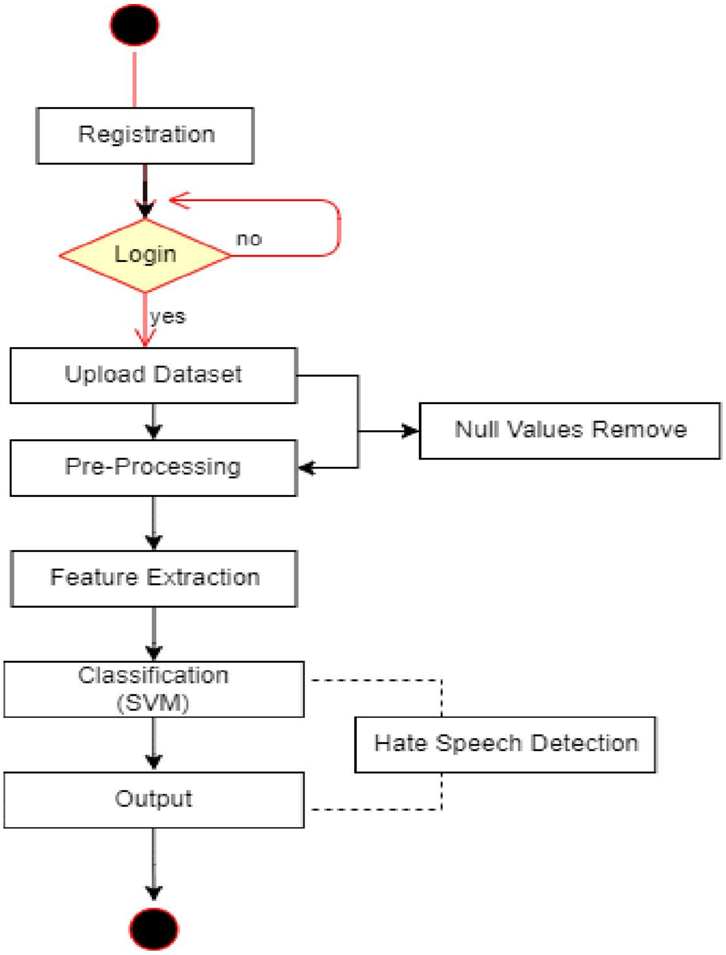

fig: Block Diagram

Hate speech and abusive language will be automatically identified, allowing the platform to detect and delete hate speech and abusive language much more rapidly and efficiently. Among the different machine learning methods, deep learning, a subset of machine learning, is widely used in Natural Language Processing (NLP) to address the problem of text classification

To simplify the process of classifying of hate speech and abusive language we have used machine learning approach to detect hate speech and abusive language from the twitter data. And To classify hate speech and abusive language from the tweets we have implemented machine learning algorithms like SVM, Logistic Regression .

IV. METHODOLOGY

This study focuses on machine learning classification models, hate speech and abusive language datasets, and the most significant features of hate speech and abusive language models in light of such a widespread phenomenon. We applied a survey and content analysis to retrieve and analyze relevant studies based on the methodology of Tranfield et al. The primary research question in this paper is to identify the dominant approaches to hate speech and abusive language and its significant datasets that are commonly used by the research community.

This section outlines the procedures followed in this SLR study to provide fair coverage of the reviewed literature. The systematic review process involved several procedures: formulation of the study questions, development of the search string, selection of the study criteria, data extraction, and data synthesis.

DOI: 10.48175/IJARSCT-15677

The goal of the study was to explore recent advances within the topic of hate speech detection and to find, evaluate, analyze, and synthesize the works conducted on the detection of hate speech on Twitter to provide a summary of all the efforts that have been achieved in the study of this subject. The strategies in Kitchenham and Charters (2007) were adopted in conducting the SLR.

V. CONCLUSION

In routine life, as the usage of social media is increased everyone seems to think like they can speak or write anything they want. Due to this thinking hate speech and abusive language has been increased so it becomes necessary to automate the process of classifying the hate speech and abusive language data. To simplify the process of classifying of hate speech and abusive language we have used machine learning approach to detect hate speech and abusive language from the twitter data. For this we have used tf-idf and bag of words methods to extract feature from the tweets. To classify hate speech and abusive language from the tweets we have implemented machine learning algorithms like SVM, Logistic Regression and Random Forest. We can conclude from the results obtained that by using Data without preprocessing and machine learning models with default parameters, Random Forest with bag of words gives best performance with 0.6580 F1 Score and 0.9629 Accuracy Score. But as explained earlier only obtaining highest accuracy is not enough when we are dealing with imbalance class dataset. For that we have used here F1 score which is quite low for data without preprocessing. To improve this we have used some preprocessing steps and gridsearch to obtained best parameter for machine learning model.Limitations of this approach is that it can be only applied to the twitter dataset so to detect hate speech and abusive language from big data can be a challenge. In future f1 score and accuracy can be improved. More machine learning techniques needs to be explored. Also different method needs to be applied to handle the imbalance class data.

REFERENCES

[1]. by RM Alhejaili • 2022 • Cited by 4 — This work provides a comprehensive view of the concepts of abusive language and hate speech. The methods used for detection will also be presented.

[2]. Abdullah Alsaeedi, Mohammad Zubair Khan, "A Study on Sentiment Analysis Techniques of Twitter Data", International Journal of Advanced Computer Science and Applications, Vol.10, No.2, 2019.

[3]. Suchita V Wawre, Sachin N Deshmukh, "Sentiment Classification using Machine Learning Techniques", International Journal of Science and Research (IJSR), Vol.6, 2015.

[4]. Ali Hasan, Sana Moin, Ahamad Karim and Shahaboddin Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts", Journal mca, 16 January 2018, Accepted 24 February 2018, Published: 27
Febryary 2018.

[5]. Vishal A. Kharde, S. S. Sonawane, "Sentiment Analysis of Twitter Data: A survey of Techniques", International Journal of Computer Applications (0975-8887) Volume 139, No.11, April 2016 .

[6]. Suchita V Wawre, Sachin N Deshmukh, "Sentimental Analysis of Movie Review using Machine Learning Algorithm with Tuned Hyperparameter", International Journal of Innovative Research in Computer and Communication Engineering, Vol.4, Issue 6, June 2016

DOI: 10.48175/IJARSCT-15677