Machine Learning Approach for Hate Speech Detection for Social Media

[1]Dashamraj Tarachand Bante, [2]Sakshi Tanaji Shinde,
[3] Tanuja Dadaji Khainar, [4]Namrata Uttam Panpatil

Student
Department of Computer Engineering,
MET's Institute of Engineering
Nashik, Maharashtra, India

***Abstract*- The social network, which is such an important part of our lives, is plagued with online impersonation and fraudulent accounts. In online social networks, fake profiles are commonly used by intruders to carry out malicious activities such as harassing a person, identity theft, and privacy violations. As a result, determining whether an account is genuine or fraudulent is one of the most difficult tasks in OSN (Online Social Network). Toxic online content has become a major issue in today's world due to an exponential increase in the use of the internet by people of different cultures and educational backgrounds. Differentiating hate speech and offensive language is a key challenge in the automatic detection of toxic text content. This propose system is to automatically classify tweets on social media into two classes: hate speech and non-hate speech using the sentimentanalysis technique. This proposed system builds using NLTK and Text blob library to detect the sentiment (positive, negative, and neutral) of a given word. Text blob trained with millions of sentences using naïve bays algorithm.**

***Key Words*: Hate Speech, Machine Learning, Encryption, Detection, Count Vectorizers, Decision Tree Classifier, Text blob, NLTK, Natural language processing.**

Published in IJIRMPS (E-ISSN: 2349-7300), Volume 11, Issue 3, May-June 2023

INTRODUCTION

Aspect-Based Sentiment Analysis (ABSA) is a type of text analysis that categorizes opinions by aspect and identifies the sentiment related to each aspect. The goal here for the ABSA system is to identify thetwo aspects – design and price – with their related sentiment. In other words, design: positive, price: negative. Aspect-Based Sentiment Analysis (ABSA) is a type of text analysis that categorizes opinions by aspect and identifies the sentiment relatedto each aspect. Using aspects, consider attributes or components of an entity (a product or a service, in our case). The sentiment polarity of a sentence is not onlydecided by its content but also has a relatively significant correlation with the targeted aspect. For this reason, propose a model by using Count Vectorizers and Decision Tree Classifier algorithm.

LITERATURE SURVEY

"Hate Speech Dataset from Turkish Tweets", Islam Mayda, Yunus Emre Demir etal., This paper explainedthat, Today, while the content produced by users on online platforms increases rapidly due to the spread ofthe internet, hate speech expressions on these platforms also increase similarly [1]. Social mediaplatforms with millions of users are especially amongthe areas where hate speech expressions are shared frequently. Popular social media companies form their own policies within the scope of combating hatespeech. However, the size of the data on the internet makes it almost impossible to do this manually. Consequently, especially in recent years, many studies have been conducted on the automatic detection of hate speech. While most of the studies inthe literature are on English, there are published studies on hate speech detection in many languages such

as German, French, Arabic, Indonesian, Portuguese. One of the main reasons for fewer studies in languages other than English is the smaller numberand size of publicly shared hate speech datasets in those languages. There is a similar situation for Turkish. Therefore, within the scope of the study, a hate speech dataset comprising 10,224 Turkish tweetswas generated and shared publicly. Tweets 4 MachineLearning Approach for Hate Speech and Offensive Language Detection for Social Media were labeled ashate, offensive, and none, and tweets tagged as hate were assigned subclass labels such as ethnic, religious, sexist, and political, which express the type of hate. In the first step of the labeling process, two annotators labeled all tweets separately. In the comparison made after this process, it was seen that the agreement rate in the given labels was 92.5. Afterwards, the two annotators discussed the tweets they gave different labels by exchanging ideas and increased the agreement rate to 98.4. For the remaining tweets, the opinion of the third evaluator was sought. After the labeling process, it was seen that the rate of hate speech in the data set was 22.8. This publicly available data set, which is a first for Turkishin terms of its scope and size, is expected to be an important resource for automatic hate.

"Multi-modal Hate Speech Detection using MachineLearning", Fariha Tahosin Boishakhi; PonkojChandra Shill etal [2]. This paper presents, With the continuous growth of internet users and media content, it is very hard to track down hateful speech in audio and video. Converting video or audio into text does not detect hate speech accurately as human sometimes uses hateful words as humorous or pleasant in sense and also uses different voice tones or show different action in the video. The state-of- the-art hate speech detection models were mostly developed on a single modality. In this research, a combined approach of multi-modal system has been proposed to detect hate speech from video contents byextracting feature images, feature values extracted from the audio, text and used machine learning and Natural language processing.

"Usage of user hate speech index for improving hatespeech detection in Twitter posts", Ehlimana Krupalija etal [3]. This paper studied that, social media isan important source of real-world data for sentiment analysis. Hate speech detection models can be trainedon data from Twitter and then utilized for content filtering and removal of posts which contain hate speech. This work proposes a new algorithm for calculating user hate speech index based on user post history. Three available datasets were merged for the purpose of acquiring Twitter posts which contained hate speech. Text preprocessing and tokenization was performed, as well as outlier removal and class balancing. The proposed algorithm was used for determining hate speech index of users who posted tweets from the dataset. The preprocessed dataset wasused for training and testing multiple machine learning models: k-means clustering without and withprincipal component analysis, naïve Bayes, decisiontree and random forest. Four different feature subsets of the dataset were used for model training andtesting. Approach for Hate Speech and Offensive Language Detection for social media to improve classification accuracy. The results showthat the usage of user hate speech index, with orwithout other user features, improves the accuracy of hate speech detection.

"A Detection of hate speech in Social media memes: Acomparative Analysis", Ajay Nayak, etal [4]. This paper presents, work projects light upon the challenges of hate speech detection in memes and demonstrates the various machine learning model to automatically detecthate in the internet memes. Memes are the visual content shared on the social media in the form of combination of picture and some textual phrases to depict light humor or jokes. However, some images inthe form of memes can also be used to convey misinformation and hate, so their early automatic detection is necessary to stop the hate spreading to widerange of users or population which may cause unrest and harm to human life and property. In this paper, thehateful meme dataset by Facebook AI has been used totest the various unimodal and a multimodal approach tobaseline performance for these models and highlight the challenges these hate memes pose to the community.

"Political Hate Speech Detection and Lexicon Building: A Study in Taiwan", Chih-Chien Wang, etal [5]. In this paper, here is the minimal restriction to users' speech in cyberspace. The Internet provides a space where people can freely present their speech, which puts a Utopian sense of freedom of speech into practice.However, the appearance of hate speech is a significantside effect of online freedom of speech. Some users use hate speech to attack others, making the attacked targets uncomfortable. The proliferation of hate speechposes severe challenges to cyber society. Users may hope that social media platforms and online communities promote anti-hate speech. However, hate speech detection is still a developing technology that requires system developers to create a method to detect unacceptable hate speech while maintaining the onlinefreedom of speech environment.

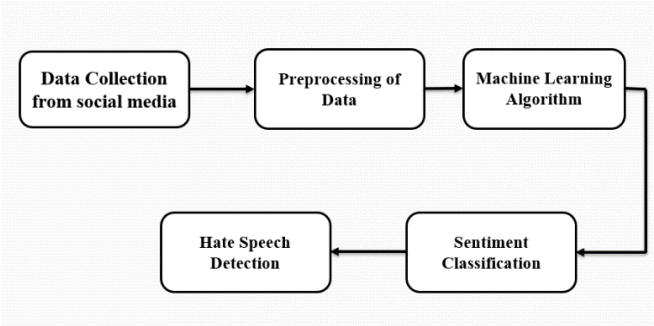
PROPOSED SYSTEM:

The detection of hate speech on social media is an important issue that many platforms are actively working to address.

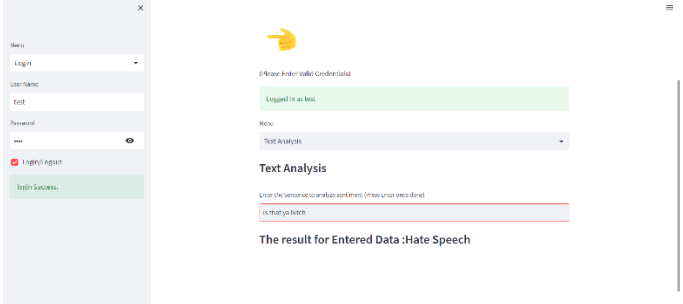
This proposed system consists of different methodologies as follows:

1. Data Collection: Collect a diverse and representative dataset of social media posts and comments, including both examples of hate speech and non-hateful content.This dataset should cover English language, regions, and demographics to ensure broad coverage.
2. Preprocessing: Clean the collected data by removing noise, irrelevant content, and personal identifiable information (PII). Tokenize the text into words or subwords, remove stop words, and perform other necessary preprocessing steps.
3. Feature Extraction: Extract relevant features from the preprocessed text data. These features can include word frequencies, n-grams, and syntactic or semantic features. Additionally, consider incorporating contextual information such as user profiles, post metadata, and historical behavior to improve the accuracy of the detection.
4. Model Training: Utilize machine learning techniquesto train a hate speech detection model. You can explore different approaches such as rule-based models, supervised learning algorithm(Count Vectorizers and Decision Tree Classifier).
5. Annotation and Labeling: Annotate the collected data with labels indicating whether each instance contains hate speech or not. This labeling process may require human annotators who are well-versed in the guidelinesand policies for hate speech detection.
6. Model Evaluation and Tuning: Split the labeled data into training and testing sets. Evaluate the performanceof the trained model on the test set using appropriate metrics such as accuracy, precision, recall, and F1score. Fine-tune the model by experimenting with different architectures, hyperparameters, and training strategies to improve its performance.
7. Integration and Deployment: Once the hate speech detection model achieves satisfactory performance, integrate it into the social media platform's infrastructure. The model can be deployed as an automated system that scans user-generated content in real-time, flagging or removing instances of hate speech.
8. Regular Updates and Maintenance: Maintain anongoing effort to update the hate speech detection system to adapt to evolving language patterns, emerging trends, and new forms of hate speech. Stay informed about the latest research in the field and incorporate advancements into the system to enhance its effectiveness.

SYSTEM ARCHITECTURE



IMPLEMENTATION AND RESULT



APPLICATIONS:

1. Social Media Platform: Hate speech detection systems can be implemented on social media platforms to automatically identify and flag offensive or discriminatory content. This helps in maintaining a safe and respectful online environment.
2. Content Moderation: Online platforms, forums, and websites can utilize hate speech detection systems to monitor user-generated content and filter out offensive or harmful language. This ensures that the content aligns with community guidelines and standards.
3. News and Media Organizations: Hate speech detection systems can be employed by news outlets and media organizations to identify and filter out hate speech in user comments or in content submitted by contributors. This helps maintain a responsible and inclusive media environment.
4. Law Enforcement and Legal Applications: Hate speech detection systems can assist law enforcement agencies and legal authorities in identifying and investigating cases of hate speech, especially in instances where it may cross legal boundaries or incite violence.
5. Research and Analysis: Hate speech detection systems can aid researchers and analysts in studying patterns and trends related to hate speech. By analyzing large amounts of data, researchers can gain insights into the prevalence, impact, and factors contributing to hate speech.

CONCLUSION

This paper presents the development and evaluation of a hate speech detection system using machine learning. The system shows promising results in accurately identifying hate speech using algorithms(Count Vectorizer and Decision tree Classifier), contributing to the efforts of combating hate speech online. However, challenges such as evolving language and the need to balance freedom of expression should be considered. Ongoing refinement, human oversight, and transparency are necessary for responsible implementation. The combination of technology, education, and dialogue is crucial in creating a more inclusive online environment.

REFERENCES:

1. NARISA ZHAO, HUAN GAO, XIN WEN, AND HUI LI, Received January 12, 2021, accepted January 15, 2021, date of publication January 19, 2021,date of current version January 27, 2021. "Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis".
2. Xingyou Wang1, Weijie Jiang2, Zhiyong Luo,Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2428–2437, Osaka, Japan, December 11-17 2016. "Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts".
3. MohammadAL-Smadi, Mahmoud M.Hammad, Sa'ad A.Al-Zboon, SajaALTawalbeh, ErikCambria, "Gated Recurrent Unit with Multilingual Universal Sentence Encoder for Arabic Aspect-Based SentimentAnalysis" Knowledge based System is an international and interdisciplinary journal in the fieldof artificial intelligence, 2021.
4. Mohsen Ghorbani, Mahdi Bahaghighat, Qin Xin and Figen Ozen, Journal of ¨ Cloud Computing: Advances, Systems and Applications (2020), LSTM, CNN: a deep learning approach for sentiment analysis in cloud computing.
5. E. Walter, Cambridge advanced Learner's dictionary, Cambridge Univ. Press, 2012.
6. W. Warner and J. Hirschberg, "Detecting HateSpeech on the World Wide Web" in , New York, NY: Columbia University, Department of Computer Science, pp. 10027D.
7. S. Macavaney, Hao-Ren Yao-E. Yang, K. Russell, N. Goharian and O. Frieder, "PLOS ONE "Hate speech detection: Challenges and solutions"",2019 Master Business Analytics Department of Mathematics Faculty of Science, August, 2018.