Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 122-126, 3, 2024
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 122-126, 3, 2024
Copyright © 2024 IJANSER
Research Article

Comparative Analysis of Machine Learning Techniques for Hate Speech Identification on Social Media

Sajida Perveen*

[1]*Department of Computer Science, National Textile University, Faisalabad, Pakistan*

Email of the corresponding author: Sajida.uaar@gmail.com

(Received:18 April 2024, Accepted: 25 April 2024)

(2nd International Conference on Scientific and Innovative Studies ICSIS 2024, April 18-19, 2024)

ATIF/REFERENCE: Perveen, S. (2024). Comparative Analysis of Machine Learning Techniques for Hate Speech Identification on Social Media. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(3), 122-126.

Abstract – Identification of hate speech on social media has become a critical challenge due to its detrimental impact on individuals and communities. Machine learning models have emerged as a potential solution to identify and mitigate hate speech. This research aims to conduct a comparative analysis among various Machine Learning (ML) techniques for hate speech identification, with the primary objective of identifying an optimal algorithmic combination that is efficient, simple, and easy to implement while yielding optimal results. Stochastic Gradient Descent (SGD), Decision tree (C4.5) and KNN models were implemented to accomplish the task. This study utilizes a labelled dataset of 49159 tweets to detect hate speech. Accuracy, precision, recall, and F1-score measures were incorporated to evaluate the models' performance, and how well these models can differentiate between instances of hate speech and those that are not. The Stochastic Gradient Descent (SGD) algorithm demonstrated remarkable accuracy (96%), precision (94%), and recall (96%) on the test dataset, highlighting its efficacy in hate speech detection compared to Decision Tree (DT) and K-Nearest Neighbors (KNN). These results pave the way for developing robust solutions, contributing to a safer and more inclusive digital environment.

Keywords – Machine Learning, Stochastic Gradient Descent (SGD), Decision Tree (C4.5), K-Nearest Neighbors (KNN), Hate Speech, Offensive Language, NLP And Social Media.

I. INTRODUCTION

Hate speech, defined as cruel, derogatory, or discriminatory language targeting groups or individuals based on their sexual orientation, gender, color, or religion, has become a detrimental issue on contemporary social media platforms [1]. The overwhelming and rapid growth of online communities and the ease of sharing information have led to the quick spread of hate speech, causing substantial harm to individuals and society as a whole [2, 3]. Hence, effective techniques are required to optimally identify and address this challenging issue, ensuring the security of internet users and fostering a more inclusive online community.

Hate speech has profound impact that extends beyond the realm of the internet [4]. Therefore, recently there has been a notable increase in efforts to combat hate speech, with scholars, legislators, and social media companies developing algorithms and strategies to identify and manage such harmful content online

[5]. Automated hate speech detection leveraging machine learning techniques has shown significant potential in this regard [6].

Nevertheless, in recent past numerous techniques have been proposed for hate speech identification [6, 7]. There is still a need for more consensus on the most effective and precise algorithmic combinations. Previous studies have delved into individual techniques such as RNN architecture [8] and SVM [9]. However, to obtain the optimal balance between precision, usability, computational efficiency and simplicity. Therefore, it's crucial to conduct a comprehensive comparative analysis of various ML algorithms [10, 11].

Therefore, this research closely examines various ML techniques for hate speech identification on social media platforms, aiming to bridge this gap. We also aimed to investigate the strengths and weaknesses of each technique to develop an optimal technique that delivers outstanding performance while remaining practical for real-world implementation.

II. MATERIALS AND METHOD

The proposed technique incorporated three widely adopted ML techniques for hate speech identification, comprising four key steps: data preprocessing, optimal feature extraction, model building and identification of Hate speech using tweets data obtained from the twitter social media platform as depicted in Figure 1. In the initial phase of data preprocessing word encoding and TF-IDF are employed to capture relevant information from the text. The subsequent phase, model learning, entails training the binary classification model on the prepared dataset, allowing it to discern between hate speech and non-hate speech based on the extracted features. Finally, the classification output is generated, providing predictions for new instances of text, thereby identifying and categorizing them as hate speech or otherwise.

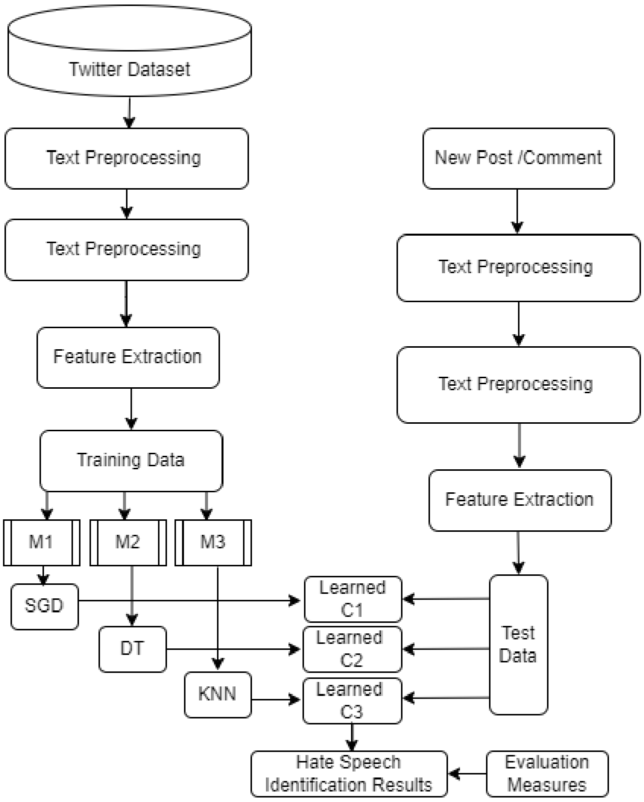


Figure 1. Flowchart of the proposed methodology

A. Dataset

The study utilizes a labelled dataset consist of tweets sampled from the Twitter social media platform. The dataset consists of a collection of speech. This dataset consists of 49,159 tweets data and broadly divided into two categories, 0 and 1 (tweets that contained non-offensive content and tweets that contained offensive content respectively). For this work, only the English tweets considered relevant in the dataset (https://thecleverprogrammer.com/2020/08/19/hate-speech-detection-model/). The URLs, hashtags and keywords prevalent in regular tweets are also present in this data. The dataset was in a CSV file with each entry containing the corresponding ID, Text and corresponding labels.

B. Data Preprocessing and Vectorization

In the initial phase of data preparation, a labeled hate speech dataset is collected and organized. Therefore, in preprocessing phase, the text dataset is cleaned to enhance its suitability for analysis. Unwanted elements such as undesired characters, URLs, numbers, and special symbols are systematically removed using regular expressions. This meticulous cleaning process ensures that the text data is in a standardized and refined format, laying the foundation for subsequent analytical steps.

C. Feature Extraction using Count Vectorizer

Following the pre-processing, the cleaned text documents are processed using the Count vectorizer. The primary purpose of the Count vectorizer is to convert the textual data into a matrix of token counts. Each document is represented as a vector, with each element corresponding to the frequency of a specific word in that document.

D. TF-IDF Transformer Transformation

The feature extraction step is pivotal, involving the transformation of raw text data into a format suitable for ML techniques. After the counting process, we incorporated the TF-IDF (Term Frequency-Inverse Document Frequency) to convert the count-based features into normalized TF-IDF features.TF-IDF assigns weights to words based on their significance within a document relative to their frequency across all documents. This normalization is crucial as it addresses the issue of common words that might be prevalent across multiple documents, ensuring that words carrying more discriminatory power are appropriately emphasized. The resulting TF-IDF vectors not only capture the frequency of words in a document but also highlight their importance in distinguishing one document from another.

E. Model Training

In the model training phase, we incorporated Stochastic Gradient Descent (SGD) classifier, K-Nearest Neighbors (KNN) and Decision Tree algorithm. The SGD classifier, known for its efficiency and scalability, iteratively updates its parameters using a stochastic approximation of the gradient descent algorithm, making it suitable for large-scale datasets [13] KNN, a non-parametric algorithm, classifies instances based on the majority class among its k nearest neighbors, making it intuitive and easy to interpret [12,13]. Decision tree, on the other hand, partition the feature space into regions based on decision rules, enabling the visualization of decision-making processes [14]. The objective of this study is to identify the most effective model while optimizing the use of features. Therefore, the dataset is further divided into training and testing datasets. The training and test dataset consists of 31,962 and17197 tweets respectively. The training dataset further consists of a total of 2242 hate tweets while the remaining tweets belong to non-hate speech. Each algorithm is trained using training dataset.

Subsequently, this study extensively evaluates the performance of the Stochastic Gradient Descent (SGD) classifier, K-Nearest Neighbors (KNN), and Decision tree algorithms on the test dataset. The model performance was assessed based on metrics such as accuracy, precision, recall, and F1-score [15]. Additionally, considerations are made for the computational complexity and interpretability of each model, providing valuable insights into their suitability for hate speech detection tasks.

III. RESULTS AND DISCUSSION

The study used a thorough methodology to assess the effectiveness of three machine learning algorithms for detecting hate speech using social media data. The evaluation phase focused on assessing the models' performance on a test dataset, utilizing metrics such as accuracy, precision, recall, and F1-score. A comparative analysis is also conducted among SGD, KNN, and Decision Tree classifiers as can be seen in Figure 2.

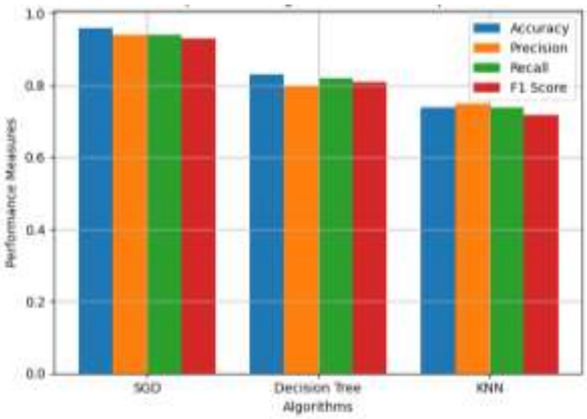
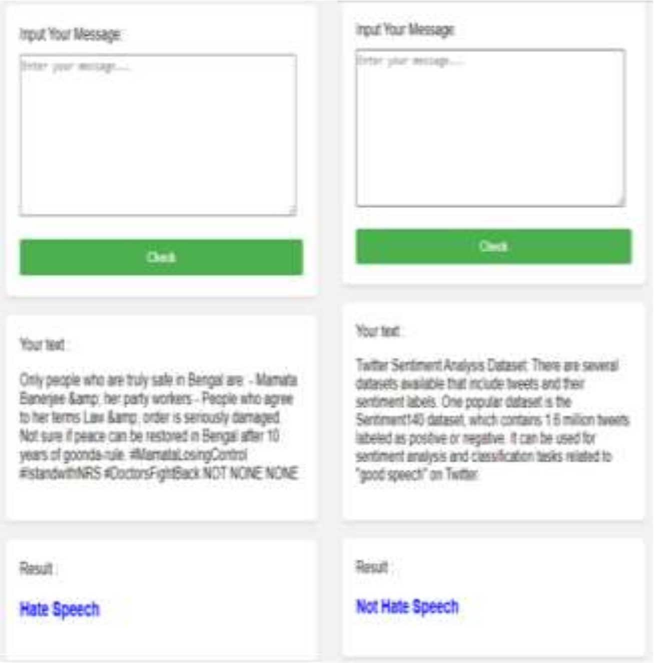


Figure 2. A comparative analysis among SGD, DT and KNN for hate speech identification using Twitter data.

Experimental results demonstrated that The SGD classifier exhibited robust performance, achieving high accuracy (96%), indicating that it effectively identifies hate speech instances with high precision (94%). Furthermore SGD Achieved comparatively high score 93%, which is a balanced measure considering both precision and recall, demonstrating overall good performance. Decision tree demonstrated competitive results, particularly in precision and recall metrics. KNN, although interpretable, showed slightly lower performance compared to the other algorithms. However, it demonstrated a precision of 75%, which is lower than both SGD and Decision Tree, implying a higher rate of false positives in identifying hate speech. Figure 3 graphically presents the results obtained from our proposed method user-friendly GUI.



Your text :

Only people who are truly safe in Bengal are - Mamata Banerjee &amp; her party workers - People who agree to her terms Law &amp; order is seriously damaged Not sure if peace can be restored in Bengal after 10 years of goonda-rule. #MamataLosingControl #IstandwithNRS #DoctorsFightBack NOT NONE NONE

Result :

Hate Speech

Your text :

Twitter Sentiment Analysis Dataset: There are several datasets available that include tweets and their sentiment labels. One popular dataset is the Sentiment140 dataset, which contains 1.6 million tweets labeled as positive or negative. It can be used for sentiment analysis and classification tasks related to "good speech" on Twitter.

Result :

Not Hate Speech

Figure 3. Results obtained from our proposed on randomly selected test data sample.

Overall, the experimental results demonstrated that the proposed research contributes valuable insights into the effectiveness and applicability of machine learning algorithms for hate speech detection tasks in social media, offering direction for future investigations and advancements in the fight against hate speech online.

IV. CONCLUSION

In this proposed study, three ML models are incorporated for identifying hate speech. SGD classifier achieved better results compared to its simplicity. In the future deep learning can be used for hat speech identification and can be evaluated on a much bigger dataset.

REFERENCES

[1] Di Fátima B, Munoriyarwa A, Gilliland A, Msughter AE, Vizcaíno-Verdú A, Gökaliler E, Capoano E, Yu H, Alikılıç İ, González-Aguilar JM, Tsene L. (2023). Hate Speech on Social Media: A Global Approach. Pontificia Universidad Católica del Ecuador.

[2] Beausoleil LE.(2019). Free, hateful, and posted: Rethinking First Amendment protection of hate speech in a social media world. BCL Rev, 60:2101.

[3] Ruwandika ND, Weerasinghe AR.(2018). Identification of hate speech in social media. In2018 18th international conference on advances in ICT for emerging regions (ICTer), (pp. 273-278). IEEE.

[4] Gagliardone I, Gal D, Alves T, Martinez G. (2015). Countering online hate speech. Unesco Publishing.

[5] Tontodimamma A, Nissi E, Sarra A, Fontanella L. (2021). Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics. 126:157-79.

[6] Fortuna P, Nunes S.(2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR). 31;51(4):1-30.

[7] Abro S, Shaikh S, Khand ZH, Zafar A, Khan S, Mujtaba G. (2020).Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications,11(8).

[8] Mazari AC, Kheddar H. (2023). Deep learning-based analysis of Algerian dialect dataset targeted hate speech, offensive language and cyberbullying. International Journal of Computing and Digital Systems.

[9] Elzayady H, Mohamed MS, Badran KM, Salama GI. (2023).A hybrid approach based on personality traits for hate speech detection in Arabic social media. International Journal of Electrical and Computer Engineering,1;13(2):1979.

[10] Vakili M, Ghamsari M, Rezaei M. (2020). Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. arXiv preprint arXiv:2001.09636.

[11] Kamal M, Bablu TA. (2022). Machine Learning Models for Predicting Click-through Rates on social media: Factors and Performance Analysis. International Journal of Applied Machine Learning and Computational Intelligence. 12(4):1-4.

[12] Akuma S, Lubem T, Adom IT. (2022) Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. International Journal of Information Technology. Dec;14(7):3629-35.

[13] Muneer A, Fati SM. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet. 29;12(11):187.

[14] Song YY, Ying LU. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 4;27(2):130.

[15] Kovatchev V, Gupta S, Das A, Lease M. (2022). Fairly Accurate: Learning Optimal Accuracy vs. Fairness Tradeoffs for Hate Speech Detection. arXiv preprint arXiv:2204.07661.