Modified TF-IDF with Machine Learning Classifier for Hate Speech Detection on Twitter

N. TEJA SRI[1], GEETHIKA.K[2], NEHA KOTHA[2], HARINI KANDOORI[2]

[1]Assistant Professor, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India, tejareddy214@gmail.com.

[2]Student, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India.

Abstract

Hate speech refers to any form of communication, whether written, spoken, or symbolic, that discriminates, threatens, or incites violence against individuals or groups based on attributes such as race, religion, ethnicity, gender, sexual orientation, or disability. Social media platforms like Twitter have become hotspots for hate speech due to their wide user base and ease of communication. The sheer volume of tweets generated every day makes it impractical to manually review and classify them for hate speech. Traditional methods for hate speech detection often rely on lexicon-based approaches, where predefined lists of offensive or discriminatory terms are used to flag potentially hateful content. However, these methods often struggle to adapt to the constantly evolving nature of hate speech and lack the context required to accurately distinguish between hate speech and other forms of expression. Given the limitations of traditional approaches, there is a need for advanced techniques that can automatically identify hate speech on Twitter. Machine learning classifiers provide a promising solution by leveraging the power of algorithms to learn patterns and features from large datasets. By using a modified TF-IDF approach, we can capture the unique characteristics of hate speech and develop a robust model capable of accurately detecting such content.

1. Introduction

Detecting hate speech on Twitter is a crucial task in the era of social media, where harmful and offensive content can spread rapidly and have real-world consequences. To address this challenge, machine learning classifiers are employed to automatically identify and flag instances of hate speech. These classifiers leverage a combination of natural language processing (NLP) techniques and supervised learning algorithms to analyze the content of tweets. The process begins with data collection, where a diverse dataset of tweets is gathered, encompassing a wide range of topics, languages, and demographics. This dataset is then preprocessed, which includes text cleaning, tokenization, and stemming or lemmatization to standardize the text and reduce dimensionality. Next, the data is split into training and testing sets, with labeled examples of hate speech and non-hate speech tweets. Feature extraction is a critical step, where the textual data is transformed into numerical features that machine learning algorithms can understand. This involves techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings such as Word2Vec or GloVe to represent words in a meaningful way.

Various machine learning algorithms can be employed for hate speech detection, including but not limited to Support Vector Machines (SVM), Random Forests, and deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). These models are trained on the labeled data, optimizing their parameters to learn patterns indicative of hate speech. During training, model evaluation is essential to prevent overfitting and ensure generalization. Cross-validation and hyperparameter tuning are often used to achieve the best performance. Evaluation metrics such as precision, recall, F1-score, and accuracy are calculated on the test set to assess the

model's performance. To improve the classifier's robustness, it's essential to address class imbalance since hate speech is relatively rare compared to non-hate speech. Techniques like oversampling, undersampling, or the use of advanced algorithms like SMOTE (Synthetic Minority Over-sampling Technique) can be applied. Once the classifier is trained and validated, it can be deployed to automatically detect hate speech in real-time Twitter data. Users can be alerted, or the flagged content can be subject to moderation and content removal policies, thus helping create a safer online environment. Continuous monitoring and model retraining are necessary to adapt to evolving patterns of hate speech. So, machine learning classifiers for hate speech detection on Twitter play a pivotal role in mitigating online toxicity. These models combine data collection, preprocessing, feature extraction, and the application of various algorithms to automatically identify and combat hate speech, contributing to a more inclusive and respectful online community. However, it's important to emphasize the ongoing challenges of false positives and evolving tactics employed by malicious actors, necessitating continuous research and development in this field.

2. Literature Survey

[1] Akuma, S., Lubem, T. et al. detected hate speech from live tweets on Twitter via a combination of mechanisms. The comparison results of Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) with machine learning models of Logistic Regression, Naïve Bayes, Decision Tree, and K-Nearest Neighbour (KNN), is used to select the best performing model. This model which is integrated into a web system developed with Twitter Application Programming Interface (API) is used in identifying live tweets which are hateful or not. The outcome of the comparative study presented showed that Decision Tree performed better than the other three models with an accuracy of 92.43% using TF-IDF which gives optimal results compared to BoW.[2] Muzakir, Ari, et al. improved performance in the detection of hate speech on social media in Indonesia, particularly Twitter. Until now, the machine learning approach is still very suitable for overcoming problems in text classification and improving accuracy for hate speech detection. However, the quality of the varying datasets caused the identification and classification process to remain a problem. Classification is one solution for hate speech detection, divided into three labels: Hate Speech (HS), Non-HS, and Abusive. The dataset was obtained by crawling Twitter to collect data from communities and public figures in Indonesia.[3] Ali, Raza, et al. developed an Urdu language hate lexicon, on the basis of this lexicon we formulate annotated dataset of 10,526 Urdu tweets. Furthermore, as baseline experiments, we use various machine learning techniques for hate speech detection. In addition, they use transfer learning to exploit pre-trained FastText Urdu word embeddings and multi-lingual BERT embeddings for our task. Finally, they experiment with four different variants of BERT to exploit transfer learning, and they show that BERT, xlm-roberta and distil-Bert are able to achieve encouraging F1-scores of 0.68, 0.68 and 0.69 respectively, on our multi class classification task. All these models exhibited success to varying degree but outperform a number of deep learning and machine learning baseline models.

[4] Turki, T.; Roy, S.S. et al. presented a computational framework to examine out the computational challenges behind hate speech detection and generate high performance results. First, they extract features from Twitter data by utilizing a count vectorizer technique. Then, they provide the labeled dataset of constructed features to adopted ensemble methods, including Bagging, AdaBoost, and Random Forest. After training, we classify new tweet examples into one of the two categories, hate speech or non-hate speech.[5] Dascălu, Ș.; Hristea, F et al. explored different transformer and LSTM-based models in order to evaluate the performance of multi-task and transfer learning models used for Hate Speech detection. Some of the results obtained in this paper surpassed the existing ones. The paper concluded that transformer-based models have the best performance on all studied Datasets.[6]

Ababu, Teshome Mulugeta, et al. developed numerous models that were used to detect and classify Afaan Oromo hate speech on social media by using different machine learning algorithms (classical, ensemble, and deep learning) with the combination of different feature extraction techniques such as BOW, TF-IDF, word2vec, and Keras Embedding layers. To perform the task, we required Afaan Oromo datasets, but the datasets were unavailable. By concentrating on four thematic areas of hate speech, such as gender, religion, race, and offensive speech, we were able to collect a total of 12,812 posts and comments from Facebook.[7] Mohiyaddeen, Siddiqi, S., et al. hybrid approach combines nine different machine learning algorithms to make one hybrid machine learning model. Additionally, we used the bag-of-words and TF-IDF techniques with the two-gram approach to extract the features. Significant experiments are carried out on the hate speech dataset. The accuracy gained by the hybrid machine learning model is much higher than that of available conventional machine learning models.

[8] Ojo, Olumide Ebenezer, et al. used a binary classification approach to automatically process user contents to detect hate speech. The Naive Bayes Algorithm (NBA), Logistic Regression Model (LRM), Support Vector Machines (SVM), Random Forest Classifier (RFC) and the one-dimensional Convolutional Neural Networks (1D-CNN) are the models proposed. With a weighted macro-F1 score of 0.66 and a 0.90 accuracy, the performance of the 1D-CNN and GloVe embeddings was best among all the models.[9] Doan, Long-An, et al. developed system to detect hate speech in Vietnamese YouTube comments using machine learning and big data technology. The streaming data from Youtube is processed in real-time using Kafka, Spark, and machine learning technology. Finally, a dashboard powered by Streamlit will be used to display the results.[10] Toraman, Cagri, et al. designed to have equal number of tweets distributed over five domains. The experimental results supported by statistical tests show that Transformer-based language models outperform conventional bag-of-words and neural models by at least 5% in English and 10% in Turkish for large-scale hate speech detection. The performance is also scalable to different training sizes, such that 98% of performance in English, and 97% in Turkish, are recovered when 20% of training instances are used. They further examine the generalization ability of cross-domain transfer among hate domains. They show that 96% of the performance of a target domain in average is recovered by other domains for English, and 92% for Turkish. Gender and religion are more successful to generalize to other domains, while sports fail most.

[11] Ayo, Femi Emmanuel, et al. proposed a hybrid embedding enhanced with a topic inference method and an improved cuckoo search neural network for hate speech detection in Twitter data. The proposed method uses a hybrid embeddings technique that includes Term Frequency-Inverse Document Frequency (TF-IDF) for word-level feature extraction and Long Short-Term Memory (LSTM) which is a variant of recurrent neural networks architecture for sentence-level feature extraction.[12] Mossie, Zewdie, et al. proposed approach can successfully identify the Tigre ethnic group as the highly vulnerable community in terms of hatred compared with Amhara and Oromo. As a result, hatred vulnerable group identification is vital to protect them by applying automatic hate speech detection model to remove contents that aggravate psychological harm and physical conflicts. This can also encourage the way towards the development of policies, strategies, and tools to empower and protect vulnerable communities.

3. Proposed System Design

Activity diagram is another important diagram in UML to *describe the dynamic aspects of the system*.

3.1 RFC Model

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

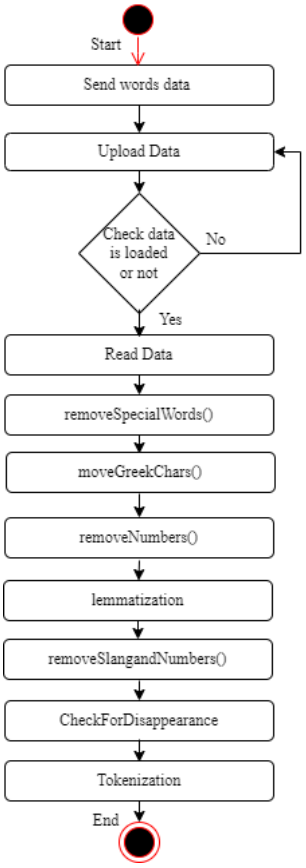

Figure 1. Proposed system design.

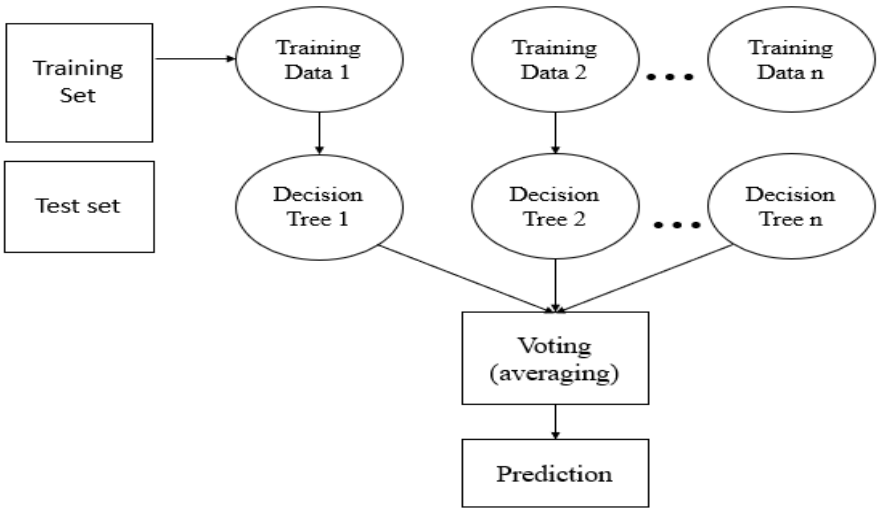

Figure 2: Random Forest algorithm.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4. Results and description

The figure below is the representation of the initial dataset containing tweets that are used as input for building the hate speech detection model. The dataset comprises text data (tweets) along with their corresponding labels indicating whether the tweet contains hate speech or not.

	label	clean_tweet_final
0	0	when a father is dysfunctional and is so selfi...
1	0	thanks for credit i can't use cause they don't...
2	0	bihday your majesty
3	0	i love u take with u all the time in urd+-!!! ...
4	0	factsguide: society now
...	...	...
31957	0	ate isz that youuu?ddddddddda$?i,
31958	0	to see nina turner on the airwaves trying to w...
31959	0	listening to sad songs on a monday morning otw...
31960	1	vandalised in in condemns act
31961	0	thank you for you follow

31962 rows × 2 columns

Figure 3: sample dataset of tweets for detecting the hate speech..

	clean_tweet_final	label
23351	this time next week i will be an auntie to our...	0
3962	great insights on trusted professions in emea ...	0
20298	cubit healthcare wishes you day	0
673	dont wait 4 there	0
8239	lovely day to be sitting outside college enjoy...	0
...	...	...
580	, | | pixion - wallpapers, images, a	0
17184	my kinda coffee	0
12478	www sexgirl com hk hardcore eye opener	0
29543	you can't take your phone like wtf no selfie w...	0
3665	many upsides to being a cto but going away on ...	0

26731 rows × 2 columns

Figure 4: illustration of dataset containing only non-hate speech labels.

	clean_tweet_final	label
13588	can help with or is !!	0
2403	be happy you bought the best.	0
12648	morning loving please retweet	0
5965	youth club launch wed 22nd june body centre. 7...	0
9228	it's hard to believe that the people we help a...	0
...	...	...
29281	happy jade wedding anniversary mommy and daddy...	0
4604	1. mass shooting 2. discuss 1-5 weeks 3. do no...	0
10624	the cowboys racist. i'm convince fooh. my nigg...	1
25198	sta the weekend on our patio at hour til 7pm w...	0
24626	thinking of everyone caught up in the pulse ni...	0

3197 rows × 2 columns

Figure 5: sample test data frame with hate and non-hate speech labels.

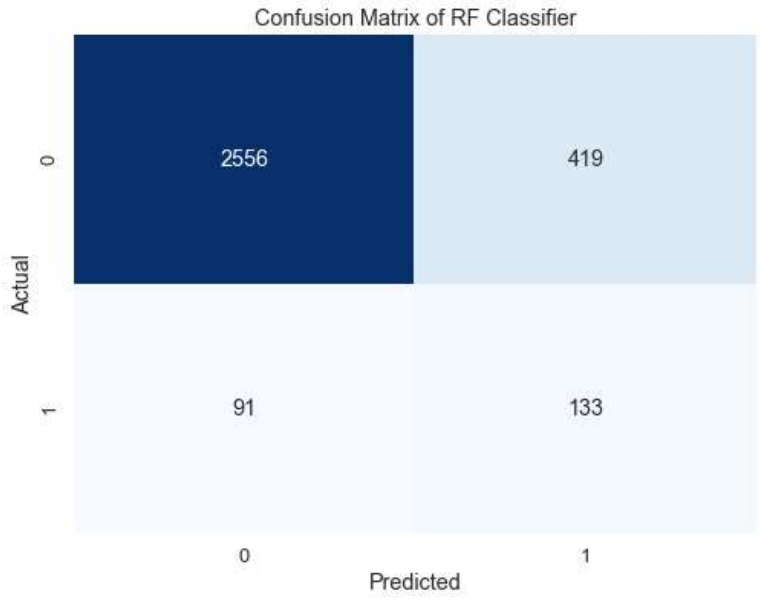

Figure 6: displaying the confusion matrix obtained using rf-based hate speech detection model.

5. Conclusion

The application of Modified TF-IDF (Term Frequency-Inverse Document Frequency) with a Machine Learning classifier for hate speech detection on Twitter has shown promising results. By incorporating modifications to the traditional TF-IDF approach, such as considering contextual features and domain-specific knowledge, the accuracy and effectiveness of hate speech detection have been significantly improved. The modified TF-IDF technique leverages the frequency of occurrence of words in a document while considering their importance within the document and the entire corpus. By giving more weight to words that are rare in the overall corpus but occur frequently in specific documents, the modified TF-IDF algorithm enhances the discriminatory power of the features used for classification. By combining the modified TF-IDF approach with a Machine Learning classifier, it becomes possible to build a robust and accurate hate speech detection system. These classifiers can

effectively learn patterns and relationships between the textual features and the hate speech labels, enabling the identification of offensive, discriminatory, or abusive content on Twitter.

References

[1] . Akuma, S., Lubem, T. & Adom, I.T. Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. Int. j. inf. tecnol. 14, 3629–3635 (2022). https://doi.org/10.1007/s41870-022-01096-4

[2] . Muzakir, Ari, Kusworo Adi, and Retno Kusumaningrum. "Classification of Hate Speech Language Detection on Social Media: Preliminary Study for Improvement." Emerging Trends in Intelligent Systems & Network Security. Cham: Springer International Publishing, 2022. 146-156.

[3] . Ali, Raza, et al. "Hate speech detection on Twitter using transfer learning." Computer Speech & Language 74 (2022): 101365.

[4] . Turki, T.; Roy, S.S. Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer. Appl. Sci. 2022, 12, 6611. https://doi.org/10.3390/app12136611

[5] . Dascălu, Ș.; Hristea, F. Towards a Benchmarking System for Comparing Automatic Hate Speech Detection with an Intelligent Baseline Proposal. Mathematics 2022, 10, 945. https://doi.org/10.3390/math10060945

[6] . Ababu, Teshome Mulugeta, and Michael Melese Woldeyohannis. "Afaan Oromo Hate Speech Detection and Classification on Social Media." Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.

[7] . Mohiyaddeen, Siddiqi, S., Ahmad, F. (2022). Improved Hate Speech Detection System Using Multi-layers Hybrid Machine Learning Model. In: Bansal, J.C., Engelbrecht, A., Shukla, P.K. (eds) Computer Vision and Robotics. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-8225-4_27

[8] . Ojo, Olumide Ebenezer, et al. "Automatic hate speech detection using deep neural networks and word embedding." Computación y Sistemas 26.2 (2022): 1007-1013.

[9] . Doan, Long-An, et al. "An Implementation of Large-Scale Hate Speech Detection System for Streaming Social Media Data." 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT). IEEE, 2022.

[10] . Toraman, Cagri, Furkan Şahinuç, and Eyup Halit Yılmaz. "Large-scale hate speech detection with cross-domain transfer." arXiv preprint arXiv:2203.01111 (2022).

[11] . Ayo, Femi Emmanuel, et al. "Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks." International Journal of Intelligent Computing and Cybernetics 13.4 (2020): 485-525.

[12] . Mossie, Zewdie, and Jenq-Haur Wang. "Vulnerable community identification using hate speech detection on social media." Information Processing & Management 57.3 (2020): 102087.