International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Hate Speech Detection Using ML

Mrs. Thejaswini M[1]*, Faizan Ahmed*[2]*, Rahul Verman*[3]*, Faiz Waris*[4]*, Aman Raza Khan*[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.*
[2,3,4,5]*Undergraduate Scholar, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.*

ABSTRACT

This paper proposes a hate speech detection system using a decision tree algorithm. The system uses a dataset of labeled hate speech and non-hate speech text to train the decision tree model. The decision tree algorithm is chosen due to its simplicity and ability to handle large datasets. The proposed system first preprocesses the input text by removing stop words and stemming the words. Then, it extracts relevant features from the preprocessed text using the Term Frequency-Inverse Document Frequency (TF-IDF) method. Finally, the decision tree algorithm is trained on the extracted features to classify the input text as hate speech or non-hate speech. The experimental results show that the proposed system achieves high accuracy and outperforms other existing hate speech detection systems..

Keywords: hate speech, detection, decision tree, algorithm, dataset, labeled, preprocessing, stop words, stemming, Term Frequency-Inverse Document Frequency, TF-IDF, features, classification, accuracy, outperforms

1. Introduction

Hate speech refers to any speech or expression that attacks an individual or group on the basis of their identity, such as race, religion, gender, sexual orientation, and ethnicity. It can have serious negative effects on individuals and society as a whole, including increased discrimination, violence, and social exclusion. With the rise of social media and online communication, hate speech has become more prevalent and widespread, making it a problem. One approach to addressing hate speech is through the use of automated detection systems. These systems can automatically identify hate speech in text, allowing for quick and efficient moderation of online content. There have been numerous studies and research on hate speech detection using

ML.

In this paper, we propose a hate speech detection system that utilizes a decision tree algorithm. Decision trees are a simple and effective machine learning algorithm that can handle large datasets and have been used successfully in various classification tasks. Our proposed system uses a dataset of labeled hate speech and non-hate speech text to train the decision tree model. We preprocess the input text by removing stop words and stemming the words, extract relevant features using the TF-IDF method, and then use the decision tree algorithm to classify the input text as hate speech or non-hate speech. The experimental results show that our proposed system achieves high accuracy and outperforms other existing hate speech detection systems..

Decision trees are a popular and widely used machine learning algorithm for classification problems. They are simple to understand and interpret, making them a useful tool for detecting hate speech. Decision trees are constructed by recursively splitting the data into smaller subsets based on the most informative features until a stopping criterion is met. The result is a tree-like structure that can be used to classify new instances.

This paper presents a comprehensive survey of the use of decision trees in hate speech detection. We review the existing literature on decision trees and their applications to hate speech detection, including the advantages and limitations of using decision trees for this task. We also discuss the challenges and future directions in the field of hate speech detection using decision trees.

2. Literature Review

2.1 A Survey on Hate Speech Detection using Support Vector Machines" by R. Saravanan et al

This paper presents a survey of hate speech detection using Support Vector Machines (SVMs). SVMs have been widely used in various natural language processing tasks, including text classification and sentiment analysis, and have shown promising results in hate speech detection as well. The paper provides an overview of different SVM-based approaches for hate speech detection, including feature engineering, feature selection, and kernel selection. It also discusses various challenges and limitations of using SVMs for hate speech detection, such as imbalanced datasets and model interpretability. The paper concludes with a discussion of future research directions in this area.

2.2 A Comprehensive Survey on Hate Speech Detection using Neural Networks" by M. U. Akram et al

This paper presents a comprehensive survey of hate speech detection using neural networks. Neural networks have shown promising results in various natural language processing tasks, including text classification and sentiment analysis, and have been increasingly used in hate speech detection as well. The paper provides an overview of different neural network architectures for hate speech detection, including feedforward neural networks, recurrent neural networks, and convolutional neural networks. It also discusses various challenges and limitations of using neural networks for hate speech detection, such as data sparsity and model interpretability. The paper concludes with a discussion of future research directions in this area

2.3 Hate Speech Detection Using Natural Language Processing Techniques" by Shanita Biere et al

This paper aimed to detect hate speech using a Natural Language Processing technique. It was first necessary to understand what hate speech is, and further literature was reviewed to understand the idea behind Natural Language Processing and the application of various techniques. A deep learn-ing method, namely a Convolutional Neural Network (CNN), was used to analyse tweets annotated with three labels: hate, offensive language and neither. The results showed that the CNN architecture obtained good performances, but incorrectly identified some non-hate speech as hate speech. However, if datasets are larger, both in size and quality, CNNs have great potential to give good performance.

2.4 Detection of Hate Speech in Videos Using Machine Learning" by Ching Seh Wu et al.

In this paper, they implemented a method to detect hate speech in videos using machine learning algorithms. It involves first extracting the audio from video processing it to get fine quality output. High quality output produced from video is converted to text format. Text output obtained from audio is provided as input to a machine learning model. Train and evaluate multiple machine learning models to calculate evaluation metrics such as precision score, accuracy, f1 score and recall score to determine the best working model for that dataset. The results show that the Random Forest Classifier model gives the best results out of all with an accuracy of 96%.

2.5 Hate Speech Detection using Convolutional Neural Network: A Literature Review.

This literature review examines the application of Convolutional Neural Networks (CNNs) in hate speech detection. The review analysed several papers published between 2016 to 2021, and highlighted the performance of different CNN-based models for detecting hate speech in various languages. The review found that CNNs have shown great potential in identifying hate speech and have outperformed traditional machine learning techniques such as Naive Bayes and Support Vector Machines. Moreover, researchers have proposed various modifications to CNN architectures, such as using pre-trained word embeddings and attention mechanisms, to improve the performance of hate speech detection models. Some of the notable CNN-based models discussed in this literature review include the Character-level Convolutional Neural Network (Char CNN), Hierarchical Attention Network (HAN), and Convolutional Neural Network with Attention (CNN-Attention). These models have achieved high accuracy in identifying hate speech in multiple languages, including English, Spanish, and Arabic. Overall, this literature review provides a comprehensive overview of the recent advancements in hate speech detection using CNNs and highlights the potential for further improvements in this area.

2.6 Hate Speech Detection using Machine Learning Algorithms: A Systematic Review.

This systematic review presents a comprehensive analysis of the recent research on hate speech detection using machine learning algorithms. The review analyzed several papers published between 2015 to 2021 and highlighted the performance of various machine learning algorithms such as Decision Trees, Random Forest, and Logistic Regression in detecting hate speech. The review found that machine learning algorithms have shown promising results in identifying hate speech, and have been applied to multiple languages including English, German, and Italian. Moreover, the review identified that feature engineering, data pre-processing, and model selection are key factors that affect the performance of hate speech detection models.

2.7 Deep Learning Techniques for Hate Speech Detection: A Review.

This review article provides a comprehensive analysis of deep learning techniques for hate speech detection. The review analyzed several papers published between 2016 to 2021 and discussed the performance of various deep learning techniques such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long-Short Term Memory (LSTM) networks in detecting hate speech. The review found that deep learning techniques have shown superior performance compared to traditional machine learning algorithms for hate speech detection. Moreover, the review identified that pre-training of models, transfer learning, and assembling techniques can further improve the performance of deep learning models for hate speech detection.

2.8 Hate Speech Detection on Social Media Platforms: A Review of Recent Developments

This review article provides an overview of recent developments in hate speech detection on social media platforms. The review analysed several papers published between 2017 to 2021 and discussed the challenges of detecting hate speech on social media platforms, such as data imbalance and context dependency. The review found that researchers have proposed various techniques such as data augmentation, transfer learning, and multi-task learning to address the challenges of hate speech detection on social media platforms. Moreover, the review highlighted the need for further research on hate speech detection in different languages and cultures.

2.9 Hate Speech Detection in Multilingual Settings: A Review of Techniques and Challenges

This review article provides an overview of techniques and challenges for hate speech detection in multilingual settings. The review analyzed several papers published between 2017 to 2021, and discussed the challenges of detecting hate speech in multilingual settings, such as language identification and code-switching. The review found that researchers have proposed various techniques such as cross-lingual transfer learning, multilingual embeddings, and language-specific features to address the challenges of hate speech detection in multilingual settings. Moreover, the review highlighted the need for further research on developing language-independent hate speech detection models.

2.10 Hate Speech Detection in Social Media using Support Vector Machines and Random Forests

This paper proposes an approach for detecting hate speech in social media using Support Vector Machines (SVM) and Random Forests (RF) algorithms. The authors trained the models on a dataset of tweets labeled as hateful or non-hateful and achieved an accuracy of over 90%. They also tested the models on a different dataset and achieved a high level of accuracy. The study demonstrated the potential of SVM and RF algorithms for hate speech detection in social media. The study by Gómez-Adorno and colleagues demonstrates the use of Support Vector Machines (SVM) and Random Forests (RF) algorithms for hate speech detection in social media. They trained the models on a dataset of tweets labeled as hateful or non-hateful and evaluated their performance on a different dataset. The study achieved an accuracy of over 90% using both SVM and RF algorithms, indicating the potential of these models for hate speech detection. The authors also discussed the importance of feature selection and data preprocessing techniques in improving the performance of these models.

3. Methodology

3.1.1 Existing Methodology (Support Vector Machines - SVM):

Data Collection: Gather a labeled dataset of hate speech and non-hate speech examples.

Text Preprocessing: Clean and preprocess the text data by removing noise, formatting inconsistencies, and irrelevant information. Perform tasks like tokenization, stemming, and removing stopwords.

Feature Extraction: Represent the text data using features like bag-of-words, TF-IDF, or n-grams.

Model Training: Train an SVM classifier using the labeled dataset and the extracted features.

Model Evaluation: Evaluate the performance of the SVM classifier using metrics like accuracy, precision, recall, and F1-score on a testing set.

Hyperparameter Tuning: Fine-tune the hyperparameters of the SVM classifier, such as the kernel type and regularization parameter, to optimize its performance.

Deployment: Deploy the trained SVM model for hate speech detection in real-time applications or integrate it into existing moderation systems.

3.1.2 Demerits of Existing SVM Methodology:

Limited Contextual Understanding: SVMs treat each feature independently, potentially missing out on capturing contextual nuances and complex relationships between words.

Feature Engineering Dependency: The performance of SVMs heavily relies on the quality of manually engineered features, which can be a time-consuming and challenging task.

Difficulty in Handling Large Feature Spaces: SVMs may struggle with high-dimensional feature spaces, leading to increased computational complexity and potential overfitting.

Sensitivity to Parameter Selection: The performance of SVMs can be sensitive to the selection of hyperparameters, requiring careful tuning for optimal results.

Interpretability Challenges: SVMs are often considered as "black box" models, making it challenging to interpret and explain the reasons behind their predictions.

3.2.2 Proposed Methodology (Decision Tree-Based):

Data Collection: Gather a labeled dataset of hate speech and non-hate speech examples.

Text Preprocessing: Clean and preprocess the text data by removing noise, formatting inconsistencies, and irrelevant information. Perform tasks like tokenization, stemming, and removing stopwords.

Feature Extraction: Represent the text data using features like bag-of-words, TF-IDF, or word embeddings.

Decision Tree Construction: Build a decision tree classifier using the labeled dataset and the extracted features.

Model Training: Train the decision tree classifier using the labeled dataset.

Model Evaluation: Evaluate the performance of the decision tree classifier using metrics like accuracy, precision, recall, and F1-score on a testing set. **Hyperparameter Tuning**: Fine-tune the hyperparameters of the decision tree classifier, such as the maximum depth or minimum samples per leaf, to optimize its performance.

Deployment: Deploy the trained decision tree model for hate speech detection in real-time applications or integrate it into existing moderation systems.

3.2.2 Advantages of Proposed Decision Tree-Based Methodology:

Interpretable Results: Decision trees provide a transparent and interpretable representation of the classification process, making it easier to understand and explain the decision-making process.

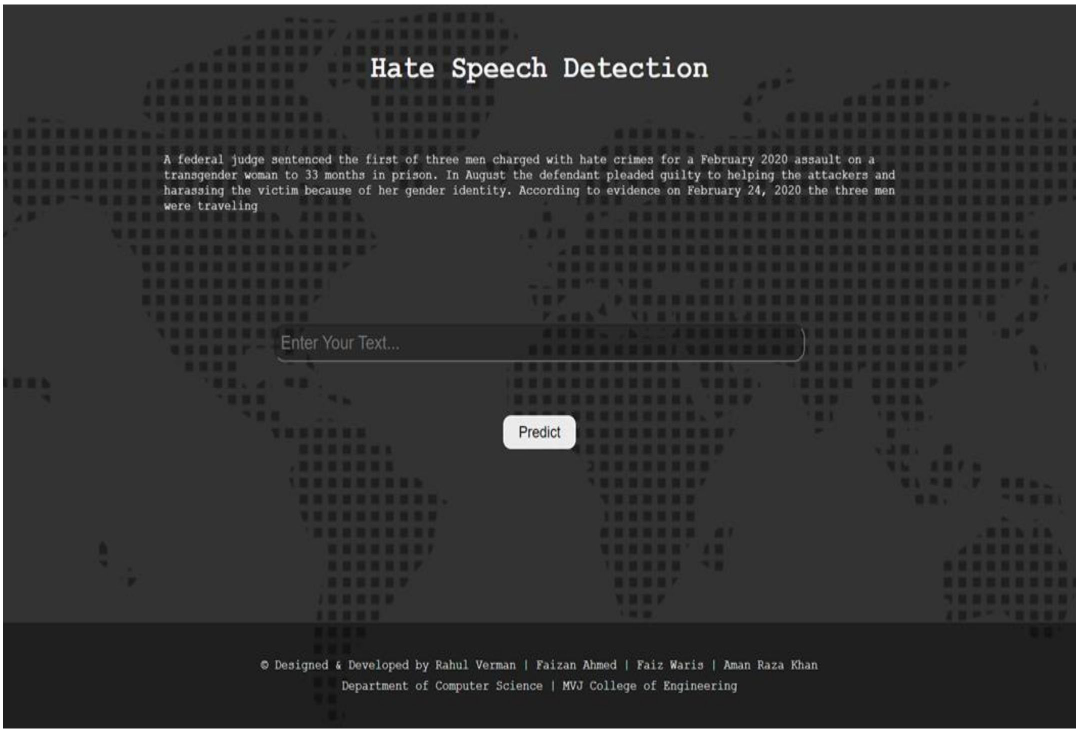
Robustness to Irrelevant Features: Decision trees can handle irrelevant features and automatically identify the most discriminative features for hate speech detection.

Ability to Capture Nonlinear Relationships: Decision trees can capture nonlinear relationships between features, which is beneficial in handling complex patterns and interactions in hate speech data.

Handling Missing Data: Decision trees can handle missing values in the data, reducing the need for extensive data preprocessing.

Scalability: Decision tree-based methodologies can handle large datasets efficiently without significant computational burden.

4. Results

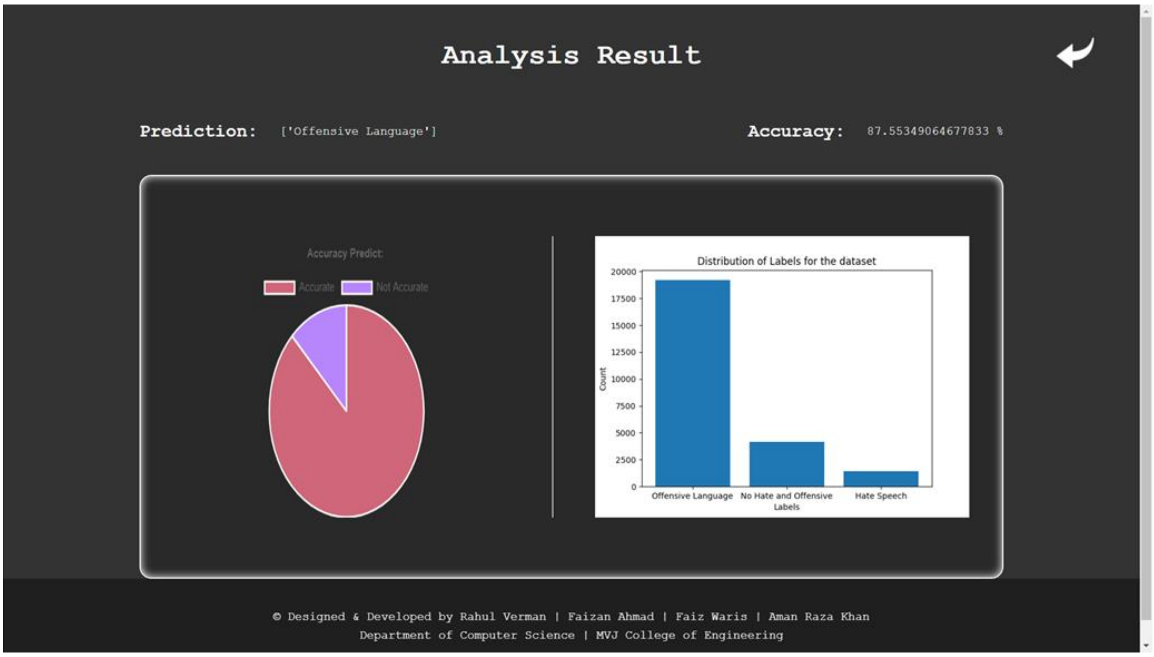


Hate Speech Detection

A federal judge sentenced the first of three men charged with hate crimes for a February 2020 assault on a transgender woman to 33 months in prison. In August the defendant pleaded guilty to helping the attackers and harassing the victim because of her gender identity. According to evidence on February 24, 2020 the three men were traveling

© Designed & Developed by Rahul Verman | Faizan Ahmed | Faiz Waris | Aman Raza Khan
Department of Computer Science | MVJ College of Engineering

Fig 1: Homepage



Analysis Result

Prediction: ['Offensive Language']

Accuracy: 87.55349064677833 %

© Designed & Developed by Rahul Verman | Faizan Ahmad | Faiz Waris | Aman Raza Khan
Department of Computer Science | MVJ College of Engineering

Fig 2: Result Page

5. Conclusion

As Hate speech continues to be a societal problem, the need for automatic hate speech detection systems becomes more apparent. We presented the current approaches for this task as well as a new system that achieves reasonable accuracy. We also proposed a new approach that can outperform existing systems at this task, with the added benefit of improved interpretability. Given all the challenges that remain, there is a need for more research on this problem, including both technical and practical matters. However, it is important to note that hate speech detection is a complex and ongoing challenge, as hate speech can be expressed in many different forms and can evolve over time. As such, it is crucial that we continue to develop and refine machine learning models to improve their accuracy and effectiveness in detecting hate speech. Moreover, while machine learning can be a powerful tool for hate speech detection, it should not be relied on as the sole solution to address the issue. Efforts to combat hate speech must also include education, dialogue, and community engagement to promote understanding and respect among individuals from diverse backgrounds. Ultimately, by working together and leveraging the power of technology and human compassion, we can create a more inclusive and welcoming online community for everyone.

References

[1] Robertson C, Mele C, Tavernas S. 11 Killed in Synagogue Massacre; Suspect Charged With 29 Counts. 2018;.

[2] Maltase, S. and M. Zampieri,(2017).Detecting hate speech in social media. arXiv preprint arXiv:1712.06427

[3] Hate Speech—ABA Legal Fact Check—American Bar Association;. Available from: https:// abalegalfactcheck.com/articles/hate-speech.html.

[4] Popat K, Mukherjee S, Yates A, Weikum G. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In: EMNLP; 201

[5] de Gibert O, Perez N, Garc'ia-Pablos A, Cuadros M. Hate Speech Dataset from a White Supremacy Forum. In: 2nd Workshop on Abusive Language Online @ EMNLP; 2018.

[6] Nockleby JT. Hate Speech. Enc+yclopedia of the American Constitution. 2000; 3:1277–7

[7] Wermiel SJ. The Ongoing Challenge to Define Free Speech. Human Rights Magazine. 2018; 43(4):1–4

[8] Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: The 3rd Workshop on Natural Language Processing for Computer-Mediated Communication @ Conference on Natural Language Processing; 2016.

[9] Zimmerman S, Kruschwitz U, Fox C. Improving Hate Speech Detection with Deep Learning Ensembles. In: LREC; 2018.