iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59053

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Hate Speech Detection

Pallerla Satwik[2], Pambala Jagan[1], Dadireddy Bhuvaneswar Reddy[3], Ms. E. Krishnaveni[4]
[1, 2, 3]*UG Student, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana*
[4]*Assistant Professor, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana*

Abstract: ***The rise in popularity of microblogging sites such as Facebook, Instagram, and Twitter has resulted in more people from different backgrounds indirectly communicating with one another. Our study aims to design an autonomous Deep Neural Network (DNN) algorithm for social media hate speech detection to tackle this problem. Using cutting-edge Natural Language Processing (NLP) techniques, the objective is to build a strong system that can recognize and categorize hate speech material in text data with accuracy. Our DNN algorithm allows for the real-time detection and moderation of offensive information, providing a proactive strategy against online hate speech. With the deployment of this technology, everyone will be able to access a safer and more welcoming online environment.***
Keywords: DNN, NLP, TensorFlow

I. INTRODUCTION

The number of people using online forums and social networks has grown exponentially in the last century. The number of active social media users as of March 2023 is 4.95 billion. Three of the most widely used micro blogging platforms for communication are Twitter, Facebook, and Instagram. Twitter generates 350000 tweets every second. People with varying backgrounds in culture and education share their opinions on daily topics such as political opinions, product reviews, movie reviews, and so forth. However, because of disagreements in viewpoint, there are instances when conversations between individuals devolve into hostile language. Bullying is one of the main causes of suicide, yet such language can take many different forms. Innocent victims of mob lynchings perish in large numbers every year. Since identifying hate speech on these kinds of websites by hand is an extremely laborious and time consuming process, this has led to the development of a model that can do so automatically. The use of such terminology must end immediately, as must its dissemination. The term "hate speech" refers to any expression that disparages an individual or group based on characteristics including race, religion, ethnicity, gender, handicap, or gender identity.

II. LITERATURE REVIEW

Due to the increasing occurrence of hate speech on social media platforms and the demand for automatic moderation solutions, researchers and practitioners have focused a great deal of emphasis on the identification of hate speech using Deep Neural Network (DNN) algorithms in recent years. In this overview of the literature, we highlight important discoveries and contributions from research on the use of DNN-based techniques for hate speech detection. Many researchers are becoming interested in sentiment analysis and opinion mining as a result of the rise in popularity of microblogging websites. A number of current publications deal with tweet classification. To address hate speech material, a variety of machine learning algorithms have been used.

In [1] Token replacement-based data augmentation methods for hate speech detection. In hate speech detection, text data from social media often lacks diversity and has too few examples of the target class, leading to overfitting. To address this, we focus on data augmentation, which involves creating synthetic samples to improve dataset quality. Our study specifically looks at token replacement, where words are substituted with synonyms. We investigate which embedding methods provide reliable synonyms, how to select words for replacement, and ensure label accuracy. Our methods, tested on two hate speech datasets with diversity issues, greatly enhance classification accuracy and offer valuable insights into token replacement techniques.

In [2] Hate Sense: Tackling Ambiguity in Hate Speech Detection Online hate speech is a big problem, but current detection methods struggle with telling apart hate speech from offensive content. This study aims to improve detection by using human-like reasoning techniques such as ontologies and fuzzy logic, along with sentiment analysis. The results show that this approach can distinguish between hateful and offensive content better than existing methods, especially when dealing with ambiguity. However, it can be challenging when there are few hateful keywords present, although the fuzzy control system helps in most cases. This research emphasizes the importance of considering the difference between hate speech and offensive content and using humanlike reasoning to improve detection.

©IJRASET: All Rights are Reserved

In [3] Automated Hate Speech Detection on Twitter. On social media like Twitter, conflicts arise due to diverse opinions, leading to a rise in hate speech. Detecting this manually is hard. Hate speech targets specific groups with aggressive language based on traits like gender or ethnicity. Our model automatically detects hate speech on Twitter using a bag of words and TFIDF approach. We trained machine learning classifiers on Twitter data and achieved 94.11% accuracy with logistic regression. This helps identify whether a tweet contains hate speech or not, making online spaces safer.

In [4] SEMAR: An Interface for Indonesian Hate Speech Detection Using Machine Learning. This research introduces SEMAR, an Indonesian hate speech detection engine using machine learning. It compares popular supervised algorithms like Naive Bayes, Decision Tree, K-Nearest Neighbors, SVM, and Logistic Regression, along with two vectorizers: Hashing and TF-IDF. SEMAR interfaces include an API and a WordPress plugin for anti-hate comment functionality. The API, implemented with SVM and TFIDF, achieved the highest accuracy (0.8707).

III. METHODOLOGIES

The aim of this study is to categorize tweets into two classes: hateful or non-hateful. The outlined steps conducted during the research are as follows. The project is designed for integration with all major social media platforms, ensuring a comprehensive approach to data analysis. This versatility allows the classification of content beyond Twitter, demonstrating the project's adaptability and broad scope.

A. Data Gathering

In the course of this project, we gathered a dataset comprising both hate speech and non-hate speech categories from Kaggle2. This dataset is structured with eight columns: id, comment text, toxic, severe toxic, obscene, and threat. To facilitate the classification process, we partitioned the dataset into two subsets. The training dataset consists of 159,572 records, while the test dataset comprises 153,165 records. The training data is specifically labeled with values of 0 and 1, where 0 denotes a non-hateful tweet, and 1 signifies a hateful tweet.

B. Data Processing

The data may contain unhelpful, missing, redundant, and inconsistent information, potentially impacting the model's accuracy and degrading its performance. This step is crucial before feeding the data into the machine learning model. This section outlines the preprocessing steps applied to tweets before the classification process. Punctuation, numbers, and special characters are eliminated, and smaller words with a length less than four, such as "pdx," "hmm," and "oh," are removed as they lack meaningful content.

Following these initial steps, the tweets are tokenized into individual words. In the final step, inflectional forms of words are eliminated, yielding the root form of each word through the use of the Porter Stemmer technique. Post- preprocessing, features are extracted to create both a bag of words and TFIDF representation. These features serve as the foundation for building a predictive model, employing a logistic regression classifier.

C. *Future Extraction*

To train a machine learning model using Tensor Flow, it's essential to provide significant features for accurate predictions. This process, called feature extraction, is crucial for training. In this study, we use tensor Flow to extract bag- of-words (BoW) and TFIDF features from tweets, serving as input for categorizing tweets into two categories. The tensor Flow-based bag-of-words method represents text by capturing word frequency in documents. Each sentence is treated as a document, creating a list of unique words with occurrences. Tensor Flow focuses on word occurrences, assuming document similarity for shared content.

While Tensor Flow’s bag-of-words highlights common words, it may lack substantial information for classification.

- It is the representation of text that represents the frequency of words within a document where each word is called token.
- Each sentence is treated as a separate document and list of all unique words in these documents are created along with their number of occurrence in all documents.

To address this, we implement the tensor Flow-based TFIDF approach, assessing word importance in a collection:

Term Frequency (TF) is a scoring of the appearance of the word in the current document. It is calculated as:

TF = (Total number of times term t appears in a document)/

(Total number oftermsin the document)

Inverse document frequency (IDF) is the scoring of how Rare the word is across the document. It is calculated as: IDF = log

(Total number of documents / Number of Documents in which term t is present)

TF-IDF = TF*ID Bag-of-Words:(BoW): xi=Count(wi,document)xi=Count(wi,document)

D. Model

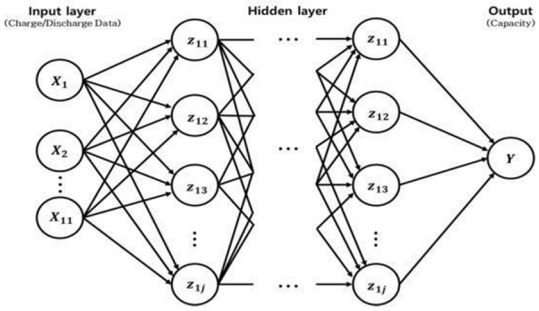
TensorFlow is a powerful open-source machine learning framework developed by the Google Brain team. It provides a comprehensive set of tools and libraries for building and training machine learning models.

TensorFlow is particularly renowned for its flexibility and scalability, allowing researchers and developers to design and deploy models across various platforms. In the context of this study, TensorFlow serves as the core engine for training the machine learning model used to classify tweets into hateful or non-hateful categories. Its intuitive design and extensive documentation make it a preferred choice, enabling efficient implementation and optimization of complex models, including the logistic regression classifier employed in this research.

E. Neural Networks

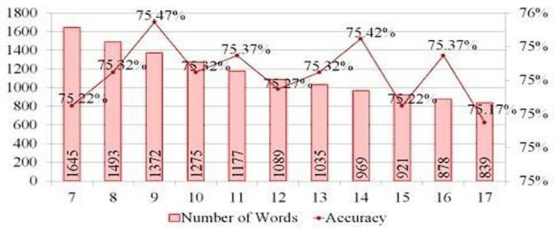
DNN stands for "Deep Neural Network," and it is a type of artificial neural network used in machine learning and deep learning.

Deep neural networks are designed to model complex patterns and representations in data, particularly when the data has multiple layers of abstraction. For the deep neural network's architecture, it is crucial to outline the number of layers, activation functions chosen (e.g., ReLU, Sigmoid), and specify the count of neurons in each layer.



F. Experimental Result

Accuracy (Correct Predictions / All predictions) of the system is defined by its f1 score which is given as F1- score=2*Precision. Recall / Precision+Recall



G. Precision

As a result, the precision formula is as follows: Precision = True positives/ (True positives + False positives) **b.Recall:** Recall = True Positive (TP) / True Positive (TP) + False Negative (FN) where, True Positive (TP) = Represents the number of positive instances correctly identified by the model.

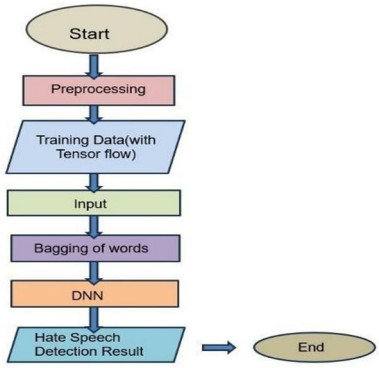


Fig 1: Block Diagram of the proposed system

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

IV. RESULTS AND DISCUSSION

In this segment, we showcase the results derived from the experiments carried out to assess the effectiveness of our Hate Speech Detection Model. We evaluate the model's performance in discerning between instances of hate speech and non-hate speech, as per the metrics outlined in the methodology. We have attained an accuracy rate of 84, which already surpasses the performance of many existing systems. Efforts are underway to further enhance this accuracy, reflecting our ongoing commitment to model improvement.

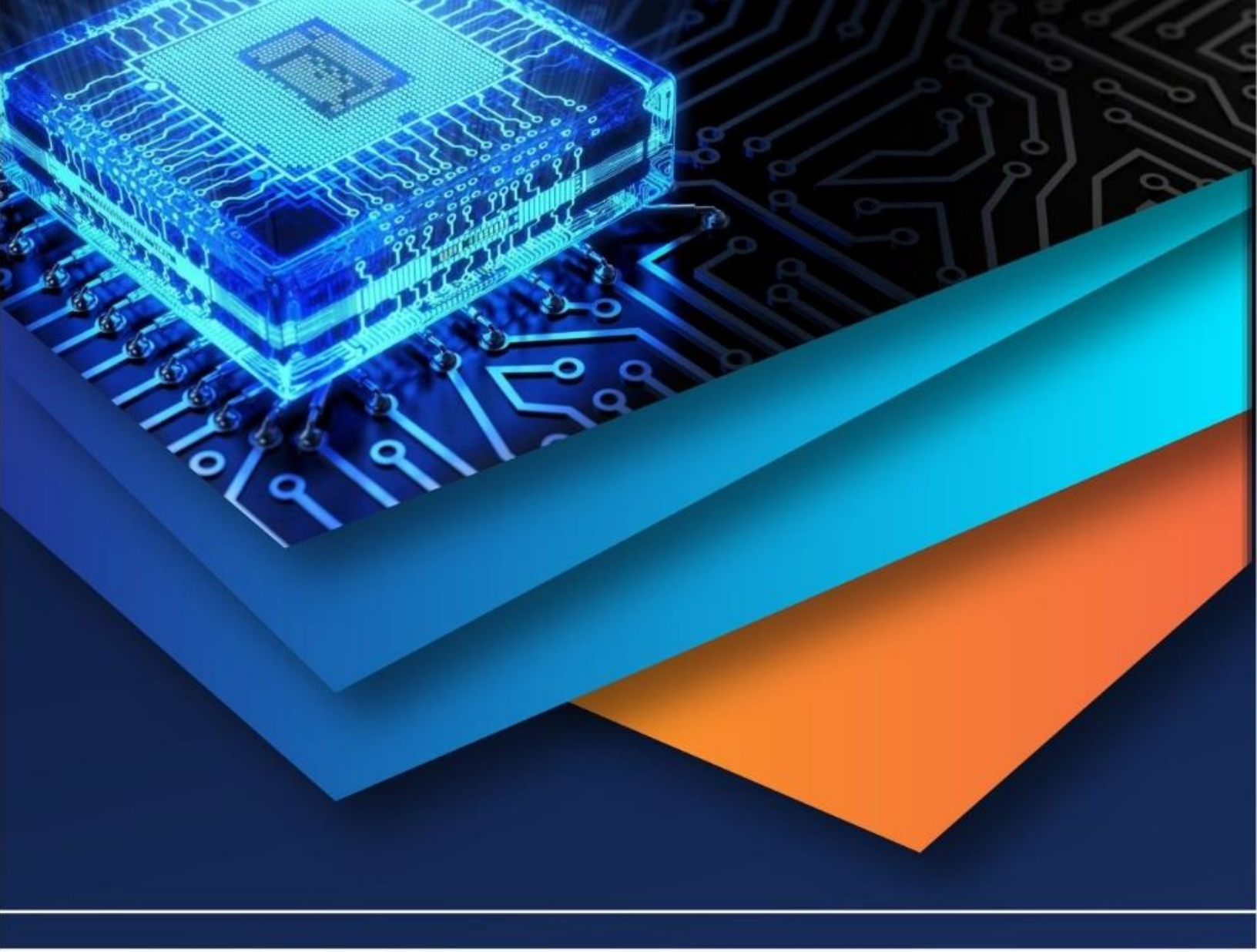
V. CONCLUSION

In this study, we introduced an approach for the automated detection of hate tweets, leveraging machine learning through the utilization of TensorFlow, NLP techniques, and a Deep Neural Network (DNN). Our methodology includes the incorporation of both a bag of words and the TFIDF approach. Contrary to the logistic regression classifier mentioned previously, our model, implemented with TensorFlow and NLP, employs a DNN architecture for tweet classification into two categories. We adopted an 80-20 split, reserving 20% of the dataset for testing and using the remaining 80% for training. With the application of TensorFlow and NLP in our model, we achieved an accuracy of 84%, reflecting its robust performance. While this accuracy is slightly lower than the bag of words and TFIDF approach, the use of deep learning techniques opens avenues for further exploration and potential improvements. Future work includes experimenting with different classifiers and incorporating additional linguistic features to enhance the model's accuracy further.toward cultivating a highperformance culture that aligns individual success with organizational triumphs.

REFERENCES

[1] Design and development of a hate speech detector in social networks based on deep learning technologies D Benito Sánchez - 2019 - oa.upm.es
[2] [HTML] Token replacement-based data augmentation methods for hate speech detection KJ Madukwe, X Gao, B Xue
[3] SEMAR: An interface for Indonesian hate speech detection using machine learning UAN Rohmawati, SW Sihwi, DE Cahyani
[4] Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe L Abualigah, YY Al-Ajlouni, MS Daoud, M Altalhi…- Social Network Analysis …, 2024 – Springer
[5] [HTML] Sarcasm detection over social media platforms using hybrid auto-encoder-based model DK Sharma, B Singh, S Agarwal, H Kim, R Sharma - Electronics, 2022 - mdpi.com
[6] A review on social spam detection: Challenges, open issues, and future directions S Rao, AK Verma, T Bhatia - Expert Systems with Applications, 2021 – Elsevier
[7] [HTML] Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends
[8] W Khan, A Daud, K Khan, S Muhammad… - … Language Processing …, 2023 – Elsevier
[9] Breaking Barriers in Sentiment Analysis and Text Emotion Detection: Toward a Unified Assessment Framework ADL Languré, M Zareei - IEEE Access, 2023 - ieeexplore.ieee.org

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)