International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Hybrid AI-Based Hate Speech Detection on Social Media Using NLP and Deep Learning

Sakshi Vilas Shinde[1], Sachin Sanjay Warude[2], Mrunal Devidas Patil[3], Isha Jagdish Patil[4], Kalpesh Sunil Pawar[5], Prof. Priti Gangurde[6]

[1,2,3,4,5]Department of Computer Science Shri Shivaji Vidya Prasarak Sanstha's, Bapusaheb Shivajirao Deore College of Engineering, Dhule, Maharashtra, India.
[6]Assistant Professor, Department of Computer Engineering, Shri Shivaji Vidhya Prasarak Sanstha's Bapusaheb Shivajirao Deore College of Engineering, Dhule (MS)Maharashtra

Abstract:
The rapid growth of social media has increased the spread of hate speech, cyberbullying, and offensive content, making manual moderation difficult. This project proposes SafeSocial AI, a hybrid hate speech detection system that uses Natural Language Processing (NLP), Machine Learning, and Deep Learning techniques to analyze both text and image-based content. TF-IDF with Support Vector Machine (SVM) and Logistic Regression models is used for text classification, while a Convolutional Neural Network (CNN) is used for image-based hate detection. The system classifies content as hate or neutral and provides intelligent moderation features such as toxicity analysis, hate-word highlighting, safe reply generation, and text sanitization. The platform is developed using Flask, HTML, CSS, JavaScript, and SQLite database integration. Experimental results show that the proposed system can effectively detect harmful social media content and help create safer online communities.

Keywords: Hate Speech Detection, Natural Language Processing, Machine Learning, Deep Learning CNN, TF-IDF, Content Moderation, Social Media Analytics, Cyberbullying.

1. Introduction

Social media platforms such as Twitter, Facebook, Instagram, and YouTube have transformed digital communication. Millions of users share opinions daily, but these platforms are also used to spread hate speech, cyberbullying, racism, sexism, and offensive language.[2]

Manual moderation of harmful content is difficult because of the enormous amount of data generated every second. Therefore, automated hate speech detection systems using Artificial Intelligence (AI) and Natural Language Processing (NLP) have gained significant attention. Traditional machine learning models provide moderate accuracy, but single classifiers often struggle with noisy and imbalanced datasets.

This research proposes a hybrid ensemble framework that combines multiple machine learning algorithms with NLP preprocessing techniques. The goal is to improve hate speech detection accuracy while reducing false positives and false negatives. The proposed framework is designed to work efficiently on multilingual and noisy social media data.

Our proposed solution involves the comparison of the performance of various machine learning algorithms about each feature set. We evaluate the performance of algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines to identify the best-performing algorithm. We conduct an in-depth analysis of the results obtained, examining the reasons for miss-classifications in our model[5].

The use of natural language processing techniques and machine learning algorithms allows us to automate the process of hate speech detection, which is crucial given the volume of content on social media platforms. Our proposed solution can be integrated into social media platforms, enabling the automatic identification and removal of hate speech content. This will enable more effective moderation and protection of users from the harmful effects of hate speech[6].

The rapid growth of social media platforms such as Instagram, Twitter, Facebook, and YouTube has transformed digital communication by allowing millions of users to share opinions, images, videos, and messages instantly. However, the increasing use of these platforms has also led to the rapid spread of hate speech, cyberbullying, racism, sexism, religious discrimination, offensive language, and harmful online content, which negatively affects individuals and communities. Manual moderation of such harmful content is difficult because social media platforms generate enormous amounts of data every second, making real-time monitoring challenging. Therefore, Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques have become essential for automatic and scalable hate speech detection. The proposed SafeSocial AI system is a hybrid web-based platform designed to detect hateful and offensive content from both text and images using TF-IDF feature extraction, Support Vector Machine (SVM), Logistic Regression, and Convolutional Neural Network (CNN) models. The system provides intelligent moderation features such as toxicity estimation, hate-word highlighting, real-time analysis, image detection, document detection, safe reply generation, and text sanitization. The platform is developed using Flask, HTML, CSS, JavaScript, and SQLite database integration to create a scalable and user-friendly content moderation system. Unlike traditional systems that focus only on text classification, the proposed framework supports multi-modal hate speech detection and aims to create safer and more responsible online communities through intelligent AI-based moderation techniques.[4]

2. Literature Review

Several researchers have proposed methods for hate speech detection using machine learning and deep learning approaches. Previous studies used techniques such as Naive Bayes, Logistic Regression, Support Vector Machines.

This study proposed a Previous studies used Support Vector Machine (SVM), Logistic Regression, and Convolutional Neural Networks (CNN) for hate speech detection. for hate speech detection on Twitter datasets. Another research paper developed a language-specific hate speech detection system for the Igbo language using neural networks and NLP techniques. These studies demonstrated the importance of hybrid AI systems for handling complex social media. data. The authors propose in [15] a methodology that uses a combination of machine learning techniques, including feature extraction, emotion recognition, and ensemble learning, to classify hate speech in social media. The results show that the proposed methodology achieves high accuracy in classifying hate speech on social media and can be used as an effective tool for detecting and preventing hate speech.
The authors propose in [7] a methodology that uses a combination of machine learning techniques, including feature extraction, emotion recognition, and ensemble learning, to classify hate speech in social media. The results show that the proposed methodology achieves high accuracy in classifying hate speech on social media and can be used as an effective tool for detecting and preventing hate speech.[6]

The authors proposed in [8] a machine learning-based approach to identify and classify hate speech in Bengali language public Facebook pages. They used a dataset consisting of 10,000 Facebook comments manually annotated as either hate speech or non-hate speech. The authors used feature engineering techniques and a Support Vector Machine (SVM) classifier to achieve an accuracy of 91.76% in detecting

hate speech. The paper concludes that their proposed approach could be an effective tool for automatically identifying and removing hateful content from Bengali language public Facebook pages.[9]

Several researchers have proposed different approaches for hate speech detection on social media using Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques. Early studies mainly focused on text classification using algorithms such as Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) to identify offensive and hateful language from social media posts and comments. Researchers also used preprocessing techniques such as tokenization, stop-word removal, stemming, lemmatization, and TF-IDF feature extraction to improve classification accuracy and reduce noise in textual data. Recent studies introduced deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and transformer-based architectures for handling complex contextual information and image-based hate speech detection. Many existing systems focus only on textual analysis and lack intelligent moderation capabilities such as toxicity analysis, real-time monitoring, and safe content generation. Some studies also highlighted challenges such as multilingual content, sarcasm detection, noisy social media text, and imbalanced datasets, which reduce model performance. To overcome these limitations, the proposed SafeSocial AI system combines NLP, Machine Learning, and Deep Learning techniques for both text-based and image-based hate speech detection. The system integrates TF-IDF, SVM, Logistic Regression, and CNN models with advanced features such as toxicity estimation, hate-word highlighting, text sanitization, document detection, and real-time moderation support to create a more scalable and intelligent social media content moderation framework.[6]

3. Proposed Methodology

The proposed SafeSocial AI system follows a hybrid methodology that combines Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques for detecting hateful and offensive content from both text and images. The methodology begins with data collection, where hate speech datasets containing harmful and neutral social media content are gathered for training and testing purposes. In the preprocessing stage, textual data undergoes cleaning operations such as lowercasing, punctuation removal, stop-word removal, tokenization, and text normalization to improve data quality and reduce noise. After preprocessing, TF-IDF (Term Frequency–Inverse Document Frequency) feature extraction is applied to convert textual content into numerical vectors that can be processed by machine learning algorithms. For text classification, machine learning models such as Support Vector Machine (SVM) and Logistic Regression are used to classify content into hate or neutral categories. For image-based hate speech detection, a Convolutional Neural Network (CNN) model is implemented to analyze uploaded images, screenshots, and hateful memes. The system also supports image URL analysis and document detection for PDF, DOCX, and TXT files. Additional intelligent moderation features such as toxicity estimation, hate-word highlighting, safe reply generation, threat severity analysis, and text sanitization are integrated to improve moderation capabilities. The frontend of the system is developed using HTML, CSS, and JavaScript, while Flask is used as the backend framework for routing and model integration. SQLite database integration is used to store user accounts, detection history, and moderation records. The overall methodology enables real-time hate speech analysis and provides a scalable framework for safer social media content moderation.[7]

1. Data Collection:
A hate speech dataset is collected from social media platforms such as Twitter and Facebook. The dataset contains hate speech, offensive language, and neutral comments.

2. Data Preprocessing:
The preprocessing stage includes:

• Tokenization
• Lowercasing
• Stop-word removal
• Punctuation removal
• Lemmatization and stemming

3. Feature Extraction:
TF-IDF (Term Frequency–Inverse Document Frequency) is used to convert textual data into numerical feature vectors. This method highlights important words while reducing the impact of common words.

4. Classification Models:
Three machine learning models are trained:
• Support Vector Machine (SVM)
• TF-IDF
• CNN (image model)

5. Deep Learning Integration:-The proposed system inte grates machine learning models for text classification and a CNN model for image-based hate speech detection.

4. System Architecture
The proposed system architecture contains the following modules:

• Input Module – Receives social media text.
• NLP Preprocessing Module – Cleans and normalizes text.
• Feature Extraction Module – Applies TF-IDF vectorization.
• Classification Module – Uses SVM and Logistic Regression for text classification and CNN fo image detection.
• Ensemble Decision Module – Combines predictions.
• Output Module – Displays classification results.

The system can classify text into three categories:
1. Hate Speech
2. Offensive Language
3. Neutral Content

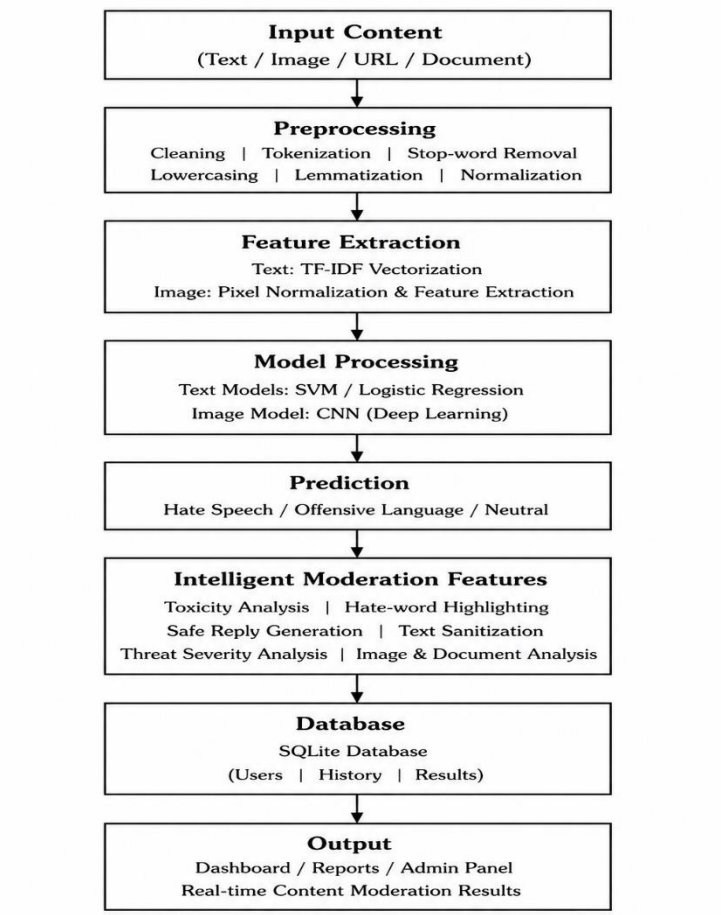

Figure 4.1: System Architecture

5. Experimental Setup

The experimental setup for the proposed SafeSocial AI system includes both hardware and software resources used for model training, testing, and web application deployment. The system was implemented using Python as the main programming language. For text-based hate speech detection, Scikit-learn was used for TF-IDF vectorization and machine learning models such as SVM and Logistic Regression. For image-based hate detection, TensorFlow and Keras were used to build and train the CNN model. Flask was used as the backend framework, while HTML, CSS, and JavaScript were used for frontend development. SQLite was used to store user details, detection history, prediction results, and admin records.[4]

The dataset was divided into training and testing sets to evaluate the performance of the models. Text data was preprocessed by removing URLs, mentions, punctuation, numbers, and extra spaces before applying TF-IDF feature extraction. Image data was resized to 128×128 pixels and normalized before being passed to the CNN model. The system was tested using different types of inputs such as text, uploaded images, image URLs, and document files. Evaluation was performed using accuracy, prediction output, toxicity level, and classification results. The complete system was executed locally using the Flask development server and tested through a web browser.

The implementation environment includes Python programming language with libraries such as:

- Scikit-learn
- Pandas

• NumPy
• NLTK
• TensorFlow

The dataset is divided into training and testing sets using a 70:30 ratio. Evaluation metrics include:
• Accuracy
• Precision
• Recall
• F1-Score
• ROC-AUC
Cross-validation techniques are used to improve reliability and avoid overfitting.

6. Results and Discussion
Experimental results indicate that the ensemble model outperforms individual classifiers. The proposed model achieved high classification .

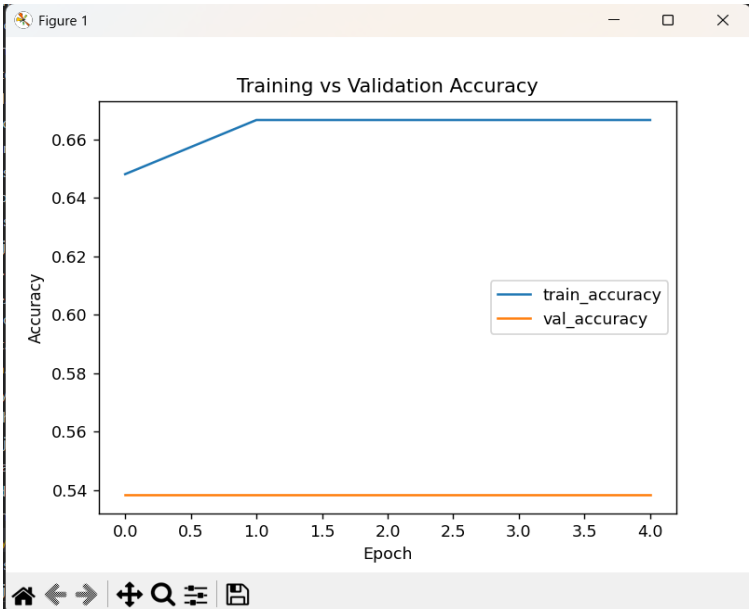

Figure: Training vs Validation Accuracy of CNN Model

The graph represents the training and validation accuracy of the CNN-based image hate speech detection model during multiple training epochs. The training accuracy gradually increased during model training, while the validation accuracy remained stable, demonstrating the learning behavior and performance consistency of the proposed deep learning model.[2]

7. Conclusion
This research presented a hybrid AI-based hate speech detection framework using NLP and ensemble machine learning techniques. The proposed system combines TF-IDF feature extraction with SVM, classifiersand through a stacking ensemble model. Exerimental analysis demonstrated that the ensemble app roach significantly improves hate speech detection. Future research can focus on multilingual hate speech detection, real-time deployment, transformer-based deep learning models, and explainable AI techniques.[6]

Overall, the results of this study suggest that ensemble learning techniques, specifically the AdaBoostClassifier algorithm, can be effective in detecting hate speech in social media. Future research can explore the use of other ensemble learning techniques and incorporate additional features such as contextual information to further improve the accuracy of hate speech detection models.[8]

REFERENCES:

[1] P. S. Br Ginting, B. Irawan, and C. Setianingsih, "Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method," in *Proc. IEEE Int. Conf. on Internet of Things and Intelligence System (IoTaIS)*, 2019.
[2] P. Ana *et al*., "An AI System for the Detection of Hate Speech Encoded in Igbo Native Language," *International Research Journal of Engineering and Technology (IRJET)*, 2024.
[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proc. AAAI Int. Conf. on Web and Social Media*, 2017.
[4] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, Jul. 2019.
[5] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in *Proc. NAACL Student Research Workshop*, 2016, pp. 88–93.
[6] P. William, R. Gade, R. Chaudhari, A. Pawar, and M. Jawale, "Machine Learning based Automatic Hate Speech Recognition System," 2022.
[7] Ariadna *et al*., "Racism, Hate Speech, and Social Media: A Systematic Review and Critique," in *Proc. Int. Conf. on Sustainable Computing and Data Communication Systems*, 2022.
[8] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Computing Surveys*, vol. 51, no. 4, Article 85, pp. 1–30, Jul. 2019, doi: 10.1145/3232676.
[9] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression and Violent Behavior*, vol. 4 0, pp. 108–118, 2018, doi: 10.1016/j.avb.2018.05.00