Hate Speech Detection of Social Media Comments in Hindi

Heet Raval
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
1032211007@mitwpu.edu.in

Roshni Singh
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
1032211185@mitwpu.edu.in

Khushi Tiwari
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
1032211126@mitwpu.edu.in

Gauransh Jain
School of Computer Engineering and Technology
MIT World Peace University
Pune, India
1032211193@mitwpu.edu.in

Prof. Laxmi Bhagwat
Department of Computer Engineering and Technology
MIT World Peace University
Pune, India
Ruchi.rani@mitwpu.edu.in

***Abstract*— The study delves into preprocessing Hindi comments for hate speech detection through natural language processing. Emotion detection is vital for understanding user sentiments, aiding recommendation systems, customer service, and market analysis. The report details step like normalization, tokenization, stop words and punctuation removal, and TF-IDF vectorization. Preprocessed data is utilized to train and assess machine learning models. Key findings underscore the efficacy of preprocessing in enhancing model performance, alongside challenges in handling Hindi text data.**

***Keywords* – Hate speech detection, Emotion detection, Hindi reviews, preprocessing, natural language processing, TF-IDF, machine learning**

1. Introduction

This paper introduces a machine learning approach for hate speech detection in Hindi comments, crucial for understanding sentiments in the digital space. Leveraging NLP techniques and supervised learning algorithms, it aims to classify comments into various emotion categories, such as hate, defamation, and more.

The research commences with thorough data collection from diverse digital sources, followed by preprocessing steps like noise removal and tokenization. Feature extraction techniques like TF-IDF vectorization are applied to represent comments as numerical features.

Multiple machine learning models are trained and evaluated, including Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, and Gradient Boosting.

This work contributes to sentiment analysis by addressing challenges specific to Hindi text, providing a reliable framework for interpreting emotions in the Hindi-speaking online community.

This paper contributes to the field of sentiment analysis by addressing the specific challenges and nuances of emotion detection in Hindi text. The proposed machine learning framework offers a reliable and efficient solution for understanding and interpreting sentiments in Hindi comments, thereby facilitating deeper insights into the emotions prevalent in the Hindi-speaking online community.

2. Literature Review

[1] Detecting Hate Speech in Hindi in Online Social Media (2023)
This paper focuses on detecting hate speech specifically in Hindi language content within online social media platforms. It discusses methodologies and approaches utilized to recognize and categorize instances of hate speech in Hindi text data, aiming to contribute to efforts aimed at combating hate speech online.

[2] Hate and Offensive Speech Detection in Hindi Twitter Corpus (2022)
This paper presents a study on hate and offensive speech detection specifically in Hindi Twitter corpus. It discusses methods and strategies employed to identify and classify hate speech and offensive language in Hindi tweets, contributing to research efforts aimed at mitigating online toxicity in social media platforms.

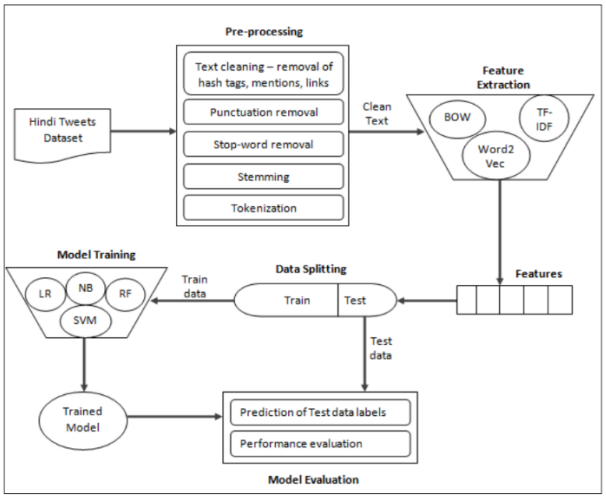
[3] Hindi-English Hate Speech Detection: Author Profiling, Debiasing,and Practical Perspectives. (2020)
This paper presents research on hate speech detection in Hindi-English code-mixed content, with a focus on author

profiling, debiasing techniques, and practical perspectives. It discusses approaches to identify hate speech in code-mixed language data and explores strategies to mitigate biases and improve the effectiveness of hate speech detection systems.

[4] Hate Speech Detection in Hindi-English Code-Mixed Social Media Text (2019)
This paper delves into the detection of hate speech in Hindi-English code-mixed social media text. It investigates the unique challenges posed by code-mixed language data and explores techniques to effectively identify and classify hate speech in such multilingual contexts, providing insights into the dynamics of online hate speech.

[5] Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media (2019)
This paper explores the task of online multilingual hate speech detection, focusing on Hindi and English social media platforms. The study investigates various techniques and approaches to detect hate speech across different languages, aiming to address the challenges posed by multilingualism in online communication.

3. System Architecture



4. Dataset Preparation

The primary objective of this study is to perform hate speech detection on Hindi comments using natural language processing (NLP) techniques. By analyzing the hate expressed in Hindi text data, the research aims to provide insights valuable for various applications such as recommendation systems, customer service, and market analysis.

4.1. Data Collection:
The data collection phase involves gathering a comprehensive dataset of Hindi reviews from various sources, including social media platforms, e-commerce websites, and forums. This diverse dataset forms the basis for subsequent analysis and model training.

4.2. Data Preprocessing:
Upon collecting the raw data, a meticulous preprocessing phase is conducted to ensure the quality and uniformity of the dataset. The dataset is then partitioned into training and testing sets to facilitate model development and evaluation.

4.3. Feature Extraction:
One primary method for feature extraction is employed to represent the text data numerically:
3.3.1. Term Frequency-Inverse Document Frequency (TF-IDF): Tf-idf Vectorizer is utilized to emphasize the importance of each word in the document corpus, providing a more nuanced feature representation.

4.4. Model Training:
Machine learning models are trained on TF-IDF features to detect hate from the text data. The models include Logistic Regression, Support Vector Machine, Naive Bayes, and ensemble methods. This step equips the system with the ability to recognize patterns and sentiments expressed in Hindi comments.

4.5. Distribution of Emotion Labels:
Model evaluation, Hyperparameter Tuning of Logistic Regression, Prediction of the label on a random comment illustrates the distribution of emotion labels in the dataset. The representation enables us to visualize the predicted label, encoded as an integer, is mapped back to its corresponding sentiment category using a reverse mapping dictionary.

5. Model Development

The workflow for hate speech detection on Hindi comments in India follows a systematic approach, encompassing key stages from data collection to model evaluation.
After training the classifiers on the preprocessed data, the models are evaluated using various performance metrics to assess their effectiveness in sentiment analysis of Hindi reviews.

5.1. Splitting Data and Model Training: Before model evaluation, the dataset is split into training and testing sets. Typically, a portion of the data (e.g., 80%) is used for training the models, while the remaining portion (e.g., 20%) is reserved for testing the models' performance. The machine learning models, including Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, and Gradient Boosting, are trained on the training dataset.

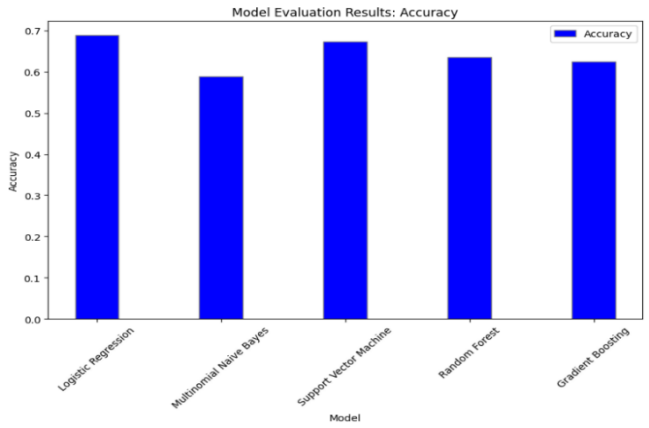
5.2. Hyperparameter Tuning: Hyperparameter tuning may be performed using techniques like grid search or random search to optimize the models' performance. This involves systematically searching for the best combination of hyperparameters (e.g., regularization parameter, kernel type) that maximize a chosen evaluation metric (e.g., accuracy).

5.3. Comparison of Models: After evaluating each model, the results are compared to determine which model performs best on the given dataset. This comparison involves analyzing accuracy scores, classification reports, and visualizations such as bar plots or ROC curves.

6. RESULTS AND DISCUSSIONS

The results reveal that preprocessing techniques significantly impact the performance of emotion detection models. Removing stopwords, punctuation, and non-Hindi characters improves the model's ability to capture meaningful patterns in the text data. Among the machine learning models evaluated, Logistic Regression achieves the highest accuracy, followed by Support Vector Machine and Random Forest. However, challenges such as class imbalance and linguistic variations in Hindi text pose limitations to model performance.

Model Evaluation Results:
The performance of various classifiers in sentiment analysis of Hindi reviews is compared using accuracy scores. The model comparison plot below illustrates the accuracy achieved by each classifier, providing insights into their relative effectiveness. The plot enables a quick visual comparison of the performance of each classifier, providing valuable insights into their relative effectiveness in classifying sentiments in Hindi reviews.



The performance of various classifiers in sentiment analysis of Hindi comments is evaluated, with Logistic Regression leading with an accuracy of 68.76%. Support Vector Machine follows closely with 67.28%. Multinomial Naive Bayes achieves a comparatively lower accuracy of 58.90%, indicating potential issues with the independence assumption. Gradient Boosting and Random Forest attain moderate accuracies of 62.48% and 63.53%, respectively. Despite Gradient Boosting's capability to handle complex relationships, it falls short in this task. Random Forest, known for robustness, lags slightly behind Logistic Regression and SVM.

Classifier	**Accuracy (%)**
Logistic Regression	68.76
Support Vector Machine	67.28
Multinomial Naive Bayes	58.90
Gradient Boosting	62.48
Random Forest	63.53

Table: Model Evaluation Results

Logistic Regression excels due to its ability to predict probabilities based on predictor variables, making it adept at distinguishing emotions in Hindi comments. Support Vector Machine's strength lies in its effectiveness in both linear and non-linear classification tasks, enabling accurate classification of emotional categories. However, Multinomial Naive Bayes struggles with lower accuracy, indicating challenges in feature independence assumption. While Gradient Boosting and Random Forest offer moderate accuracies, further optimization or feature engineering could enhance their performance.

In summary, Logistic Regression and Support Vector Machine emerge as top performers, shedding light on effective algorithms for sentiment analysis in Hindi text data. Hyperparameter tuning was performed, resulting in an increased accuracy of Logistic Regression to 70.54%. These findings offer valuable insights for businesses, governments, and social media platforms aiming to understand sentiments expressed in Hindi comments.

7. CONCLUSION AND FUTURE SCOPE

In conclusion, this report demonstrates the importance of preprocessing techniques in improving the accuracy of hate speech detection models for Hindi comments. The findings underscore the need for tailored preprocessing approaches to handle linguistic intricacies and cultural nuances inherent in Hindi text data. Future research directions include exploring advanced deep learning architectures, incorporating domain-specific lexicons, and integrating multimodal features for more comprehensive hate speech detection in Hindi comments. The insights gained from this study can inform the development of effective emotion detection systems to enhance user engagement and satisfaction in various applications.

8. REFERENCES

[1] A. Sharma and R. Kaushal, "Detecting Hate Speech in Hindi in Online Social Media," in Proc. of the 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2023, pp. 1-5, doi: 10.1109/ICCT56969.2023.10075749.

[2] Jadhav, Ishali & Kanade, Aditi & Waghmare, Vishesh & Chaudhari, Deptii.. Hate and Offensive Speech Detection in Hindi Twitter Corpus.

[3] S. Chopra, R. Sawhney, P. Mathur, and R.R. Shah, "Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives," in Proc. of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, pp. 386-393, 2020, doi: 10.1609/aaai.v34i01.5374.

[4] K. Sreelakshmi, B. Premjith, and K.P. Soman, "Detection of Hate Speech Text in Hindi-English Code-mixed Data," Procedia Computer Science, vol. 171, pp. 737-744, 2020, doi: 10.1016/j.procs.2020.04.080.

[5] N. Vashistha and A. Zubiaga, "Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media," Information, vol. 12, no. 1, p. 5, 2021, doi: 10.3390/info12010005.