Algorithm Alliance@LT-EDI-2024: Caste and Migration Hate Speech Detection

Saisandeep Sangeetham, Shreyamanisha C Vinay, Kavin Rajan G, Abishna A & B Bharathi
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Tamil Nadu, India
saisandeep2210495@ssn.edu.in
shreyamanisha2210857@ssn.edu.in
kavinrajan2210227@ssn.edu.in
abishna2210385@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Caste and Migration speech refers to the use of language that distinguishes the offense, violence, and distress on their social, caste, and migration status. Here, caste hate speech targets the imbalance of an individual's social status and focuses mainly on the degradation of their caste group. While the migration hate speech imposes the differences in nationality, culture, and individual status. These speeches are meant to affront the social status of these people. To detect this hate in the speech, our task on Caste and Migration Hate Speech Detection has been created which classifies human speech into genuine or stimulate categories. For this task, we used multiple classification models such as the train test split model to split the dataset into train and test data, Logistic regression, Support Vector Machine, MLP (Multilayer Perceptron) classifier, Random Forest classifier, KNN classifier, and Decision tree classification. Among these models, The SVM gave the highest macro average F1 score of 0.77 and the average accuracy for these models is around 0.75.

1 Introduction

In the age of rapid globalization and digital interconnectedness, social media platforms have become powerful tools for communication and community engagement. However, this unprecedented accessibility has also given rise to a darker aspect of online discourse – the proliferation of hate speech. Of particular concern is the manifestation of hate speech related to caste and migration issues, which not only perpetuates discrimination but also poses a significant threat to social harmony. As our world embraces the Digital Age, technology plays a pivotal role in connecting people through platforms like Facebook and Twitter (Drus and Khalid, 2019).

Despite its positive aspects, social media harbors drawbacks, with users sometimes engaging in discouragement or targeted hate speech. Detrimental speech on these platforms has a lasting psychological impact on victims (Gongane et al., 2022). This study highlights the surge in hate speech on social media, fuelled by anonymity and the absence of stringent controls, particularly targeting religion, gender, and race. Online communities offer insights into understanding and combating online hate speech, suggesting new dimensions for future research (Nazmine and Khan Tareen, 2021).

Social media platforms struggle to manage the constant flood of comments and posts, making it challenging to effectively monitor and control content due to the sheer volume. Finding a balance between limiting excessive posts and preserving freedom of speech poses a significant predicament. Additionally, the diverse user base, representing various backgrounds, cultures, and beliefs, further complicates the issue, contributing to the widespread problem of hate speech. (Al-Hassan, 2019).

The paper's structure includes a literature review in Section 2, task and data description in Section 3, methodology in Section 4, results and analysis in Section 5, and a conclusion in Section 6.

2 Related Works

Numerous studies have explored hate speech detection, including those focused on caste and migration (Kim et al., 2018). Davidson et al. emphasized the subjective biases in hate speech classification, highlighting the need for objective methodologies. In caste-based hate speech detection, Malmasi and Zampieri addressed challenges using lexical properties like n-grams, character n-grams, word embeddings, and paragraph embeddings (Kim et al., 2018).

Research on migration-related hate speech includes traditional and deep learning-based hate speech classification methods proposed by (Subramanian et al.,

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 254–258
March 21, 2024 ©2024 Association for Computational Linguistics

2023). Sanguinetti et al. conducted automatic hate speech detection research, creating datasets annotated with hate labels and related dimensions (Jahan and Oussalah, 2023). The overview of the hope speech detection task is given in (Kumaresan et al., 2023).

In sentiment analysis, (Vijayakumar et al., 2022) used the transformer model ALBERT for hope speech detection in multiple languages like English, Tamil, Kannada, etc. (Chakravarthi et al., 2020) proposed a Convolutional Neural Network (CNN) model outperforming traditional models for hope-speech detection. The authors of(Balouchzahi et al., 2022) performed binary and multi-class hope-speech classification. The binary task involved only two labels whereas the multi-class task involved three labels.

In the paper, (Velankar et al., 2021) used HASOC 2021 Hindi and Marathi hate speech datasets for algorithm comparison. Marathi uses binary labels; Hindi has both binary and detailed labels. Transformer models excelled, and basic models with fastText embeddings showed competitive performance. Intriguingly, after standard hyper-parameter tuning, basic models outperformed BERT-based models, especially on the fine-grained Hindi dataset.

The authors of (Roy et al., 2022) examined code-mixed language use on social media, focusing on Hindi-English, Tamil-English, Malayalam-English, Telugu-English, etc. They proposed a weighted ensemble model combining transformer-based BERT models and a deep neural network for offensive and hate speech detection. Experimental results showed the framework outperformed state-of-the-art models, achieving 0.802 and 0.933 weighted F1 scores for Malayalam and Tamil code-mixed datasets.

The authors of (Saumya and Mishra, 2021) used LSTM, deep learning, and hybrid models on Tamil and Malayalam datasets. In the paper (Ghanghor et al., 2021) applied transformer models like m-BERTcased and XLM-RoBERTa for hope speech detection, with m-BERT-cased achieving the highest F1-score. The top model for the English dataset was the 2-parallel CNN-LSTM using GloVe and Word2Vec embeddings, while the 3-parallel Bi-LSTM excelled on the Malayalam dataset.

In recent years, there's been a rise in studies addressing hate speech targeting specific groups, like caste and migration status. In today's digital age, hate speech based on caste or migration has become a significant concern. These studies showcase versatile models for sentiment analysis on social media comments. To enhance text classification accuracy, we opted for traditional models alongside a basic transformer model based on the literature survey.

3 Task and Data Description

The overview paper for this task is explained in (Rajiakodi et al., 2024). The shared Task on Caste and Migration Hate Speech detection at LT-EDI-EACL 2024 is intended to determine whether the speech text format was legitimate or imposed hate towards Caste and Migration. The dataset consists of two fields namely speech text and a label. Here, the Label indicates the above-mentioned category, and it is represented in hate and non-hate speech. The training dataset consists of around 5,355 text-converted speeches out of which 3,303 instances were labelled as non-hate speech and 2,052 instances were labelled as hate speech. The Development dataset consisted of 945 instances out of which 594 instances were labelled as non-hate speeches and 351 instances were labelled as hate speeches. Here, we used 1576 test data instances for testing the model.

4 Methodology

Several machine learning approaches may be used to achieve this task, but we chose the most effective one for the classification problems, i.e., detection of hate speech related to caste and migration.

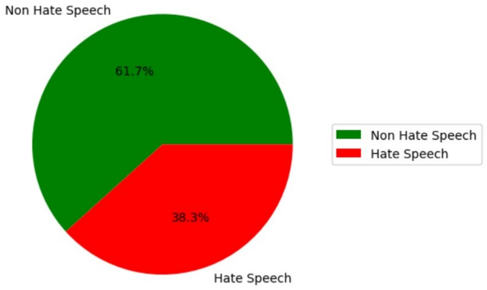

Figure 1: Data distribution in datasets

Label	Train Instances	Dev Instances
Non Hate speech(0)	3303	594
Hate speech(1)	2053	351

Table 1: Description of the Data Distribution

As shown in Figure 1, the distribution of data in the datasets indicates that 38.3% of collected data contains hate speech. Table 1 describes the data distribution of hate speech among the training and development instances.

4.1 Data Preprocessing and Cleaning

Data cleaning procedures were the first step for getting the raw data ready for use in any of the models in machine learning.

The raw data usually consists of many punctuation marks, emojis, and multiple spaces which would affect the performance of the model, hence, to ensure the uniformity of the Data, we are considering the elimination

of these. Using the popular libraries of Python such as the “Demoji” for removing all the emoji’s in the dataset, and “re” for removing the special characters, symbols, and multiple spaces in the datasets. This comprehensive pipeline of data preparation and cleaning establishes the foundation that supports subsequent phases of our research, creating a conducive environment for machine learning models to function well.
The uniform and standardized, feature-rich dataset makes the model easier to extract valuable patterns and insights, which improves the model’s overall performance.

4.2 Text Tokenization

We addressed the challenge of text vectorization by converting the raw data into a numerical format that could be utilized for a machine-learning model. Initially, we used the popular library “IndicNLP” tokenizer for tokenizing the Tamil language text to clean text. Then we transformed the entire text data into numerical vectors by utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. Therefore TF-IDF vectorization offers an accurate depiction of the text data by encoding the meaning of words in context. Specifically, we limited the feature space for the (TF-IDF) to a maximum of 5000 features. This methodological choice tries to achieve a balance between computational efficiency and the retention of essential information. This forms the foundation for the subsequent application of machine learning models in our research.

4.3 Model Selection

Selecting an appropriate machine learning model is essential, therefore our main goal is to build a model that can deal with various linguistic nuances that are present in hate speech. While still maintaining high accuracy and good classification abilities. So we chose the best suitable algorithm for this task such as by implementing some of the popular classifications such as Logistic Regression, Support Vector Machine(SVM), Multi-Layer Perceptron(MLP), Random Forest Classifier(RFC), Decision Tree, KNN.

5 Results and Analysis

5.1 Performance Metrics

In the field of Machine learning, it is critical to get the predictive model’s performance in need to determine its efficiency and suitability for practical uses. Here We determine our model performance by considering metrics such as accuracy, F1-Score, recall, precision, etc. These function as a crucial benchmark for our model.

1) Accuracy is defined as the ratio of the correctly predicted instances to the total number of instances in a dataset. It acts as a straightforward for the model’s correctness.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

2) Precision is the ability of a classification model in which it is not to label irrelevant instances as positive in normal terms it is defined as the ratio of true positives to the sum of true and false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

3) Recall which is also called sensitivity or true positive rate is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

4) F1-Score is defined as the harmonic mean of the precision and recall. It provides a balanced measure that considers both the false positives and false negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

5.2 Results and Observation

For this task, we investigated the involved application of several machine learning algorithms such as Logistic regression, Support Vector Machines (SVM), Random Forest Classifier, Decision Tree, KNN, and Multi-Layer Perceptron (MLP). Our main aim is to improve the efficiency of the models in automatically classifying the texts that are related to the cast/migration-related hate speech.

5.2.1 Comparative Model Accuracies

By evaluating the performance of various machine learning models on the given datasets. We observed the distinct accuracies across the classifiers. Logistic Regression which we achieved an accuracy of 0.711, surpassing this Support Vector Machines (SVM) outperformed this, exhibiting better discriminate power with an accuracy of 0.797, Random Forest classifier came in close to second by achieving an accuracy of 0.793, The Multi-Layer Perceptron (MLP) exhibited the competitive accuracy at 0.737, suggesting its capacity to capture sophisticated relationships within the textual data. Decision Tree achieved an accuracy of 0.746, showcasing its robustness in discerning hate speech nuances. Unfortunately, given the accuracy of 0.6402, KNN might not be performing at its best.

Tables 2, 3, and 4 show the classification reports for SVM, RFC, and Decision Tree models on the test data, respectively. Figure 2 illustrates the F1-Accuracy scores of different models.

	Precision	Recall	F1 Score	Support
0	0.80	0.91	0.85	594
1	0.80	0.61	0.69	351
Accuracy			0.80	945
Macro Avg	0.80	0.76	0.77	945
Weighted Avg	0.80	0.80	0.79	945

Table 2: Classification Report for SVM on Test Data

	Precision	Recall	F1 Score	Support
0	0.79	0.92	0.85	594
1	0.82	0.57	0.67	351
Accuracy			0.79	945
Macro Avg	0.80	0.75	0.76	945
Weighted Avg	0.80	0.79	0.78	945

Table 3: Classification Report for RFC on Test Data

	Precision	Recall	F1 Score	Support
0	0.79	0.81	0.80	594
1	0.66	0.64	0.65	351
Accuracy			0.77	945
Macro Avg	0.73	0.73	0.73	945
Weighted Avg	0.74	0.75	0.75	945

Table 4: Classification Report for Decision Tree on Test Data

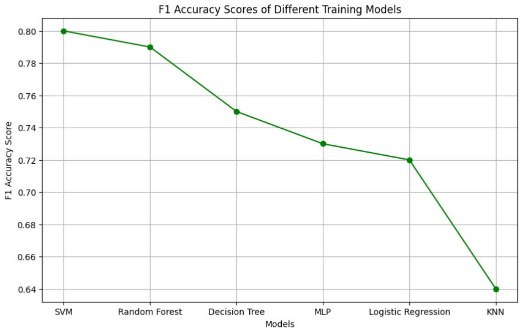

Figure 2: : F1-Accuracy Scores of Different models

6 Limitations

Our research on hate speech detection using SVM and other ML models has shown promise, but it also has notable limitations. The biased training data may not fully represent real-world instances, which challenges the models' ability to generalize. Moreover, subjective hate speech labeling introduces inconsistencies, which affects the reliability of the data.

Another limitation is class imbalance, where hate speech instances are outnumbered by non-hate speech instances, making it difficult to accurately identify and potentially leading to misclassifications. Additionally, linguistic complexity further complicates detection, as SVM and other ML models may struggle with nuances such as sarcasm, irony, and cultural references that are common in hate speech.

Furthermore, SVM models heavily rely on feature engineering, which limits the selection of features that robustly represent diverse hate speech characteristics. The "black box" nature of SVM models also raises concerns about explainability, making it difficult to interpret predictions.

To overcome these limitations, exploring innovative solutions such as improved feature engineering, diverse training datasets, and interpretable ML models is crucial. These steps will enhance the reliability of hate speech detection systems, urging future research to address these challenges.

7 Ethics Statement

“Avoid harm” our model only detects hate speech but doesn’t mentally and physically affect anyone. “Be fair and take action not to discriminate”. Equality for all and no discrimination on any grounds was done while detecting hate speech. We create opportunities for members of the organization or group to grow as professionals and for team growth.

8 Conclusion

In conclusion, we applied supervised learning models such as Random Forest, SVM, and Logistic regression to investigate hate speech identification and migration speech, with a macro F1 score of 0.77, the SVM model stood out and demonstrated its efficiency by classifying the hate speech in these specific contexts. The following research could investigate the integration of deep learning models to boost accuracy. While emphasizing the ongoing need for adaptive and more flexible classification to deal with the evolving dynamics of these conversations.

References

Hmood Al-Hassan, Areej Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. pages 83–100.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2022. Polyhope: Two-level hope speech detection from tweets.

Bharathi Chakravarthi, Vigneshwaran Muralidaran, Ruba Asoka Chakravarthi, and John McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text.

Zulfadzli Drus and Haliyana Khalid. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics.

Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing.

Geewook Kim, Kazuki Fukui, and Hidetoshi Shimodaira. 2018. Word-like character n-gram embedding. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 148–152, Brussels, Belgium. Association for Computational Linguistics.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Subalalitha Cn, Miguel Angel García-Cumbreras, Salud Maria Jimenez Zafra, José Antonio García-Díaz, Rafael Valencia-García, Momchil Hardalov, Ivan Koychev, Preslav Nakov, Daniel García-Baena, and Kishore Kumar Ponnusamy. 2023. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *LTEDI*.

Manan Nazmine and Hannan Khan Tareen. 2021. Hate speech and social media: A systematic review. *Turkish Online Journal of Qualitative Inquiry*, 12:5285–5294.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Pradeep Bhawal Roy, Snehaan Cn, and Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

Sunil Saumya and Ankit Kumar Mishra. 2021. IIIT_DWD@LT-EDI-EACL2021: Hope speech detection in YouTube multilingual comments. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 107–113, Kyiv. Association for Computational Linguistics.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, and G. Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi. 2021. Hate and offensive speech detection in hindi and marathi.

Praveenkumar Vijayakumar, S Prathyush, P Aravind, Angel Suseelan, Rajalakshmi Sivanaiah, Sakaya Milton Rajendram, and T. T. Mirnalinee. 2022. Ssn_armm@ lt-edi -acl2022: Hope speech detection for equality, diversity, and inclusion using albert model. In *LTEDI*.