SSN-Nova@LT-EDI 2024: POS Tagging, Boosting Techniques and Voting Classifiers for Caste And Migration Hate Speech Detection

A Ankitha Reddy, Ann Maria Thomas, Pranav Moorthi & B. Bharathi
Department of CSE
Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
ankithareddy2210178@ssn.edu.in, annthomas2210391@ssn.edu.in
pranav2210176@ssn.edu.in, bharathib@ssn.edu.in

Abstract

This paper presents our submission for the shared task on Caste and Migration Hate Speech Detection: LT-EDI@EACL 2024[1]. This text classification task aims to foster the creation of models capable of identifying hate speech related to caste and migration. The dataset comprises social media comments, and the goal is to categorize them into negative and positive sentiments. Our approach explores back-translation for data augmentation to address sparse datasets in low-resource Dravidian languages. While Part-of-Speech (POS) tagging is valuable in natural language processing, our work highlights its ineffectiveness in Dravidian languages, with model performance drastically reducing from 0.73 to 0.67 on application. In analyzing boosting and ensemble methods, the voting classifier with traditional models outperforms others and the boosting techniques, underscoring the efficacy of simpler models on low-resource data despite augmentation.

1 Introduction

The deep-seated phenomenon of caste discrimination in India has endured over time, with recent advancements reflecting breakthroughs in challenging these deeply ingrained biases. Despite contemporary endeavors to disentangle from the shackles of caste-based prejudices, the phenomenon still persists, exerting influence on diverse facets of individual lives (Vaid, 2014).

In the era of expanding social media platforms, marked by attributes like user anonymity, widespread accessibility, and the fostering of online communities and discourse, the identification and surveillance of hate speech rooted in caste discrimination pose a significant societal challenge. While machine learning models for hate speech detection have made significant strides in the Western context, (Corazza et al., 2020) there is a glaring gap when it comes to adapting these models to the nuanced dynamics of casteism in India. Casteism, a concept uniquely embedded in the social fabric of South Asian communities, introduces complexities that are not adequately addressed by current research and detection mechanisms. Unlike hate speech patterns prevalent in the West, caste-based discrimination in India operates within a distinct socio-cultural context, marked by intricate layers of subtext and nuanced contextual variations (Jahan and Oussalah, 2023).

The dearth of research tailored to this phenomenon unique to the Indian subcontinent hinders the effectiveness of existing models in capturing the intricacies of this societal issue, especially in the sphere of social media. It is crucial to acknowledge that the linguistic, cultural, and historical dimensions of casteism necessitate a more nuanced approach to hate speech detection, one that transcends the limitations of generic models designed for Western contexts (Sambasivan et al., 2021).

Our paper is structured as follows - Section 2 explores other publications pertaining to text classification tasks in low resource languages, Section 3 provides an analysis of the distribution of the dataset, Section 4 highlights the methodology undertaken for our proposed model and Section 5 analyses the performance metrics of the solutions and provides a conclusion.

2 Related Work

Though work has been done in text classification for low-resource languages in the recent past, it is apparent that the lack of annotated datasets has continually limited the scope of research in the field, with (Rajiakodi et al., 2024) making notable strides in this regard. This inherent drawback has severely affected the applications of widely adopted methods, including POS tagging, on morphological learning in Dravidian languages (Moeller et al.,

[1] https://codalab.lisn.upsaclay.fr/competitions/16089

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 233–237
March 21, 2024 ©2024 Association for Computational Linguistics

2021; Kann et al., 2020). Hence, data augmentation, with an emphasis on backtracking, poses as an attractive solution to aid in combating such data issues (Pingle et al., 2023; Shleifer, 2019).

With respect to classifiers utilised, research has been focused on transformer and deep learning models (Roy et al., 2022; Dowlagar and Mamidi, 2021). However, little light has been shed on the efficacy of ensemble approaches with traditional machine learning models (Kumar et al., EasyChair, 2021; Nimmi and Janet, 2021), which have proved to outperform state-of-the-art technology that requires large quantities of annotated data (Jauhiainen et al., 2021), a vision that remains to elude research in Dravidian languages.

3 Dataset Analysis

The labels given for the data were "Caste/Migration Hate Speech" and "Non-Caste/Migration Hate Speech". The data distribution is provided below in Table 1.

Category	Count
Non - Caste/ Migration Hate Speech	3,303
Caste/ Migration Hate Speech	2,052

Table 1: Data distribution

Notably, there exists a significant imbalance in the distribution of labels. This disparity may potentially hinder the implementation of our models. To rectify this imbalance and enhance the operational efficiency of our model, we implemented data augmentation on the datasets. Further details on this process will be elaborated in-depth in Section 4.

4 Methodology

4.1 Data Augmentation

Back translation stands as a data augmentation method employed in natural language processing to expand datasets. This technique involves translating a given text into another language and then back to the original language, introducing diversity and variability into the dataset.

In our proposed model, the text data was translated to English, and then translated back into Tamil as seen in Figure 1. The language was detected through the LanguageIdentifier model which is adept at discerning the language of the text, in our case, Tanglish or Romanized Tamil. This is done by computing the count of Tamil accented vowels and consonants, surpassing a predefined threshold to ascertain the text's manifestation in Romanized Tamil form. Once the source language was detected, the text was translated into the identified destination language and back into the original language. This translation was executed using the Googletrans library which implements the Google Ajax API[2]. This allowed the creation of texts that remained semantically congruent, yet diverged discernibly from its original form.

id	text	label	lang	augment_text
244	### I use to respect tamilians a lost but the way they r doing killing North Indian people has really hurted me all have life family u all kill inccent people 😔 😡 💔	0	en	### I use to respect the Tamils, but the way they killed the people of North India really hurt me. Life family is family.
5294	இவர்கள் சொல்வது எல்லாம் உண்மையான காரணமில்லை முதலாளிகள் தான் காரணம் காலை ஆறுமணியில் இரவுஒன்பதுமணி வரை கேள்வி கேட்காமல்உழைப்பதால் அவர்களைத்தான் தேடுகிறார்கள்.	0	ta	காலையில் ஆறு மணி வரை அவர்கள் கேள்வி கேட்கப்படாததால் அவர்கள் சொல்வதை முதலாளிகள் ஒரு உண்மையான காரணம் அல்ல.
592	ஜாதி பெருமையை பேசிக்கொண்டு இருப்பார்கள் இவர்கள் வந்தால் தான் இவர்களை அடக்க முடியும் சரிதான் வணக்கம் ஒரு அடிச்சு துவைக்க போறாங்க வடக்கின்ஸ் மோசமான ஆட்கள்	1	ta	சாதியின் மகிமை.அவர்கள் வந்தால், அவர்கள் அவர்களை அடக்க முடியும்.
5192	வட மாநிலத்தவனை கட்டுப்படுத்த விட்டால் தமிழ்நாட்டு மக்கள் பல விளைவுகளை சந்திக்க நேரிடும்.	0	ta	நீங்கள் வடக்கு மாநிலத்தை கட்டுப்படுத்தினால், தமிழ்நாட்டின் மக்களுக்கு பல முடிவுகள் கிடைக்கும்.
1097	it's nothing wrong people travel to earn money but in same time native people also need work hard for better life...lucky Brother you know hindi to communicate to Vadakans...Nice review	0	en	People travel to make money, but at the same time you have to work hard for the best life for the native people . Lucky brothers, you know Hindi to communicate with the Vadaka . Good Review

Figure 1: Augmentation of data using backtranslation

By creating new instances of text with similar meanings but different linguistic expressions, back translation significantly increases dataset size as seen in Table 2. The process aims to preserve semantic meaning while varying the phrasing, word choice, and sentence structure. This augmented dataset with diverse linguistic patterns should theoretically contribute to more robust model training, mitigating overfitting risks, and ultimately enhancing the performance of natural language processing models.

Category	Count
Non - Caste/ Migration Hate Speech	4,121
Caste/ Migration Hate Speech	2,834

Table 2: Data distribution after augmentation

4.2 Preprocessing

The maximization of model efficiency and the influence on performance metrics hinge significantly on data preprocessing. This fundamental process involves several key steps. Initially, the conversion

[2] https://support.google.com/code/topic/10021

of text to lowercase and the expansion of contractions promote a uniform analytical approach. Subsequently, stemming reduces words to their root form, aiding tasks such as sentiment analysis by consolidating related words. Following this, the removal of stop words expedites processing. Lastly, the removal of special characters, symbols, and emojis streamlines the text, reducing the volume for subsequent model processing.

4.3 Feature Extraction

TF-IDF, or Term Frequency-Inverse Document Frequency, is a technique for creating features from text data by measuring the importance of words in a collection of documents. It assigns higher importance to words exclusive to a small set of documents. The TF-IDF vectorizer matches each feature to a numerical value calculated from its TF-IDF score, obtained by multiplying term frequency and inverse document frequency.

4.4 POS Tagging

Linguistically, words can be categorized into various parts of speech based on their grammatical attributes. Part-of-Speech (POS) tagging is the process of assigning specific word classes to individual words in a given text. These designated tags play a crucial role in enabling models to discern the significance of different elements of speech within the provided text, thereby enhancing the model's ability to identify and comprehend the key components of speech.

4.5 XG Boost

XGBoost, a gradient boosting technique that particularly excels in the realm of structured data, employs parallel tree boosting to achieve heightened efficiency. Utilizing the weighted quantile sketch algorithm, XGBoost addresses datasets with a substantial number of zero values. The algorithm, recognized for its scalability, implements boosting as the process of minimizing a convex loss function within a convex set of functions. Regularization, incorporating both L1 and L2 regularization techniques, mitigates the risk of overfitting, while the parallel tree approach facilitates seamless scalability on clusters, reducing memory usage.

4.6 Adaptive Boosting

Adaptive Boosting (AdaBoost) strategically amalgamates multiple weak classifiers to construct a robust classifier. Employing a greedy algorithm, AdaBoost optimizes weights for each weak classifier and utilizes decision stumps to amalgamate decisions from individual classifiers. Each weak learner corresponds to a vector in an n-dimensional space, with the objective of reaching a target point where the loss function is minimized. The training process assigns weights to samples, equating to the error on the sample at the iteration point. The overarching goal is to systematically diminish the training error for the weak classifiers.

4.7 Voting Classifier

A Voting Classifier is an ensemble learning method that combines predictions from multiple individual models to make a final prediction as shown in Figure 2. It aggregates the outcomes through methods like majority voting (Hard Voting) or averaging predicted probabilities (Soft Voting). In Hard Voting, the final prediction is determined by the majority of individual classifier predictions, while in Soft Voting, the average predicted probabilities across all classifiers contribute to the final decision. These approaches aim to enhance overall model performance by leveraging the strengths of diverse base classifiers (Bartlett et al., 1998).

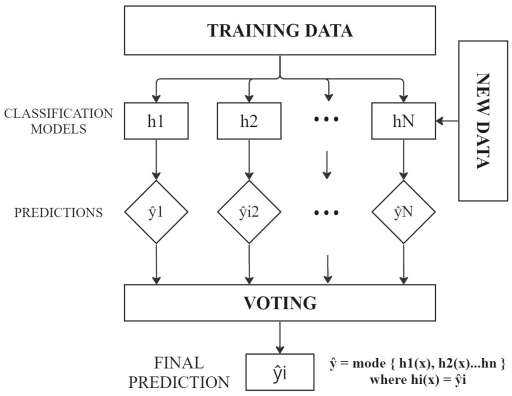

Figure 2: Structure of Voting Classifier

In this study, the ensemble approach employed soft voting, as it capitalises on the complementary strengths of the individual models, allowing for a more nuanced and robust decision-making process. Our preliminary analysis focused on identifying traditional models that boasted superior performance characteristics. After extensive experimentation, we determined that the optimal classifiers were Support Vector Machine, Random Forest, and Multinomial Naive Bayes. Each of these classifiers was incorporated into a separate pipeline along with the TfidfVectorizer. This approach ensured that the

text data underwent consistent processing across all models, ensuring consistency in the ensemble predictions.

5 Results and Analysis

The impact of the POS tagging was elementally validated by juxtaposing the performance of the SVM classifier. Analysis found that the macro-average F1 score of the model significantly decreased with the implementation of POS tagging from 0.73 to 0.67. This could be attributed to the morphosyntactic intricacies of Dravidian languages. This discrepancy stems from the profound dissimilarity in grammatical structures and semantics inherent to Dravidian languages, diverging significantly from the syntactic patterns prevalent in the Latin alphabet. Nonetheless, POS tagging heavily relies on annotated corpora for discerning patterns and relationships between words and their corresponding POS tags.

Hence, the model's ability to generalise effectively could be hindered due to the limited dataset, lack of grammar and semantic standardisation paired with the significant number of out-of-vocabulary words that may not be adequately covered in training data.

The augmentation of data, ostensibly believed to augment the efficacy of the model, yielded only a marginal enhancement in model performance, manifesting as a modest 1-2 percent improvement. Our hypothesis posits that the inherent simplicity of the operative models acts as a constraining factor, impeding their capacity to effectively leverage the augmented dataset. Nonetheless, despite marginal improvements, the augmented dataset was systematically employed for subsequent exploration and analysis.

Model	Dataset	Augmented Dataset
XGBoost	0.49	0.50
Voting Classifier	0.75	0.77
AdaBoost	0.56	0.58

Table 3: Macro-average F1- score of the proposed system using prior to and post data augmentation

The evaluation of the task is done based on the following performance metrics: Precision, Recall and macro-average F1- score.

Model	Precision	Recall	F1-Score
XGBoost	0.67	0.55	0.50
Voting Classifier	0.81	0.76	0.77
AdaBoost	0.63	0.59	0.58

Table 4: Performance of the proposed system using development data in Tamil code-mixed text

With regard to the models implemented, the superior performance of the voting classifier implies that the ensemble of traditional ML models, when combined through voting, leverages the strengths of individual models and mitigates their weaknesses.

Additionally, the AdaBoost classifier outperformed the XGBoost which may be due to the fact that AdaBoost builds a sequence of weak learners, adjusting their importance based on the errors of the previous learners; thereby enabling advantageous outcomes with low-resource languages due to its interpretability and simplicity. On the other hand, XGBoost uses a more complex and sophisticated algorithm that includes regularisation terms, parallel computation, and tree-pruning strategies.

6 Conclusion

Our approach aimed to leverage data augmentation through back translation to address the issue of sparse datasets in low-resource Dravidian languages. However, the implementation did not yield signifcant improvements in model performance.

Tangentially, Part-of-Speech (POS) tagging is exceptionally valuable in natural language processing, providing crucial insights into the grammatical structure of sentences, enabling accurate syntactic analysis, and facilitating downstream tasks like sentiment analysis and machine translation. Despite having such crucial applications, POS tagging remains ineffective on Dravidian languages, highlighting the exigency for nuanced linguistic models attuned to the unique intricacies of non-Latin script languages.

On analysis of different boosting and ensemble methods, the voting classifier incorporating traditional models proved to outperform the other models, highlighting the efficacy of simpler models on low-resource data despite data augmentation. On probing deeper, it was found that between the XGBoost and AdaBoost as well, the simpler of the two models proved to perform significantly better.

References

Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Trans. Internet Technol.*, 20(2).

Suman Dowlagar and Radhika Mamidi. 2021. EDIOne@LT-EDI-EACL2021: Pre-trained transformers with convolutional neural networks for hope speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 86–91, Kyiv. Association for Computational Linguistics.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Comparing Approaches to Dravidian Language Identification.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly Supervised POS Taggers Perform Poorly on Truly Low-Resource Languages.

S R Mithun Kumar, Nihal Reddy, Aruna Malapati, and Lov Kumar. EasyChair, 2021. An ensemble model for sentiment classification on code-mixed data in Dravidian Languages. EasyChair Preprint no. 7266.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To POS tag or not to POS tag: The impact of POS tags on morphological learning in low-resource settings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 966–978, Online. Association for Computational Linguistics.

K Nimmi and B Janet. 2021. Voting ensemble model based Malayalam-English sentiment analysis on code-mixed data.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, Geetanjali Kale, and Raviraj Joshi. 2023. Robust Sentiment Analysis for Low Resource languages Using Data Augmentation Approaches: A Case Study in Marathi. *arXiv e-prints*, page arXiv:2310.00734.

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 Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond.

Sam Shleifer. 2019. Low Resource Text Classification with ULMFit and Backtranslation.

Divya Vaid. 2014. Caste in Contemporary India: Flexibility and Persistence. *Annual Review of Sociology*, 40(1):391–410.