Transformers@LT-EDI-EACL2024: Caste and Migration Hate Speech Detection in Tamil Using Ensembling on Transformers

Kriti Singhal[1], Jatin Bedi[2]
Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology, India
[1]kritisinghal711@gmail.com, [2]jatin.bedi@thapar.edu

Abstract

In recent years, there has been a persistent focus on developing systems that can automatically identify the hate speech content circulating on diverse social media platforms. This paper describes the team "Transformers" submission to the Caste and Migration Hate Speech Detection in Tamil shared task by LT-EDI 2024 workshop at EACL 2024. We used an ensemble approach in the shared task, combining various transformer-based pre-trained models using majority voting. The best macro average F1-score achieved was 0.82. We secured the 1st rank in the Caste and Migration Hate Speech in Tamil shared task.

1 Introduction

Hate speech can be defined as the use of aggressive, abusive or threatening expressions or phrases. With the advancement of the technological age, everyone has access to the internet to voice their opinions to a large audience. However, some people may misuse this power to spread hate against a certain individual or a group of individuals based on certain distinguishing characteristics. This could be through posts on social media, blogs, videos, or comments on various platforms. Hence, it has become crucial to regulate the comments on social media platforms to avoid hurting sentiments. The shared task[1] organized by LT-EDI aimed to detect Caste and Migration hate speech in Tamil text (Rajiakodi et al., 2024).

Social media platforms have taken the freedom of speech and expression beyond global borders. Platforms like Twitter, Instagram, and YouTube allow ideas shared from one corner of the world to reach millions of people across the world in just a few milliseconds (Shanmugavadivel et al., 2022). However, the increased anonymity provided by such platforms has lead users to exploit this power by sharing opinions and ideas targeted against an individual or a group. This makes it crucial to regulate the hateful content shared online automatically to attenuate the societal harm it can cause.

In the targeted Hate Speech identification domain, Natural Language Processing (NLP) has experienced major breakthroughs in the past few years. Recent developments include Long Short Term Memory (Hochreiter and Schmidhuber, 1997), and Gated Recurrent Units (Chung et al., 2014). But with the introduction of transformers (Vaswani et al., 2017), the results have seen a paradigm shift.

Tamil is one of the twenty-two scheduled languages in the Constitution of India. Tamil is also a member of the Dravidian languages' family (Chakravarthi and Raja, 2020), which dates back over 4,500 years. However, Tamil continues to be under-resourced (Ghanghor et al., 2021). Multiple NLP approaches have also been devised with a special focus on the Indian context, this includes, IndicBERT and MuRIL (Khanuja et al., 2021).

The aim of this shared task was to build an automatic classification system which could classify whether the given text in Tamil contains caste and migration hate speech or not. In this context, the current work presents a novel approach based on transformers to classify whether a text has caste and migration hate speech in Tamil.

2 Related Work

With the recent boom in the number of internet users, many researchers worldwide have directed their efforts towards finding whether text online contains hate speech. The methodologies have evolved from the traditional machine learning models to the recent transformer-based approaches.

In the work done by Shanmugavadivel et al. (2023), a machine learning-based approach was

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

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 249–253
March 21, 2024 ©2024 Association for Computational Linguistics

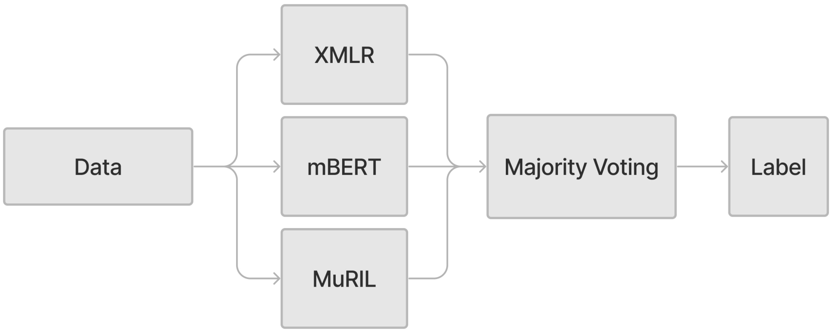

Figure 1: Proposed Methodology

proposed for the detection of abusive comments in the Tamil language after exploring various deep learning and transformer techniques. They achieved a macro-F1 score of 0.35. The work also showed how the traditional machine learning models can outperform certain deep learning and transformer-based techniques when the dataset is not large enough and complex where the deep learning approaches excel.

Similarly, Subramanian et al. (2022) experimented with multiple traditional machine learning models and transfer learning approaches. They found that even though machine learning models performed well, transfer learning approaches outperformed them. Among the different approaches that were tested, XLM-RoBERTa (Large) gave the highest accuracy. They attributed the reason to the fact that XLM-RoBERTa (Large) has more layers. Hence, the number of trainable parameters is approximately three times the other alternatives.

Bhawal et al. (2021) also observed that transformer based approaches consistently performed better on both Tamil and Malayalam text. Various models, including Logistic Regression, Support Vector Machine, etc., were implemented on the Tamil dataset, and the highest F1-score achieved was 0.82. A simple deep neural network was implemented on the same dataset, and it achieved an F1-Score of 0.89 on the Tamil text. The MuRIL model, however, outperformed both of these techniques and achieved an F1-Score of 0.91.

An ensemble approach was adopted by Roy et al. (2022). Initially, many traditional machine learning models and transformer based models were implemented individually. However, it was found that the individual models had a high misclassifictaion rate. Hence, in order to improve the accuracy, a combination of any three of the high scoring models were used for ensembling. Two different approaches were considered to ensemble the models. The first approach involved averaging the outcomes of the models and the second approach involved using custom weights which were determined by grid search method for each member of the ensemble model.

Table 1: Dataset Distribution for Caste and Migration Hate Speech Detection Task

Dataset	Label		Total
	0	1	
Dev	594	351	945
Train	3,303	2,052	5,355

3 Dataset Description

The dataset was provided by the organizers of the competition (Chakravarthi, 2020, 2022; Chakravarthi et al., 2022). The train and dev dataset is comprised of three fields, namely, id, text, and label. The test set comprised of only id and text. The labels for the dataset were 0 or 1, where 0 denoted absence and 1 denoted presence of no caste and migration hate speech, respectively. The distributions of the dev and train datasets have been shown in Table 1.

4 Methodology

Text classification is one of the most prominent tasks in NLP. It can be defined as the segregation

Table 2: Model Performance Comparison

Model	Before Pre-processing		After Pre-processing	
	F1 Score	Accuracy	F1 Score	Accuracy
MuRIL cased	0.60	0.62	0.60	0.61
XLM RoBERTa Large	0.38	0.62	0.61	0.65
Multilingual DistilBERT Base cased	0.28	0.38	0.57	0.65
XLM RoBERTa Base	0.61	0.65	0.61	0.62
Multilingual BERT Base cased	0.59	0.62	0.59	0.60
Indic BERT	0.60	0.60	0.52	0.62

of texts into different classes. Various models using a variety of word representations have been introduced in the past to tackle the text classification problem. Many of these models were based on the transformer architecture and have been pretrained on large corpora of text and made available for solving problems, including text classification. These models perform tokenization using their own tokenizers and vocabularies. However, the corpora of text these models are trained on are limited to high-resourced languages like English. Hence, this issue was solved using cross-lingual transfer learning. The proposed methodology in this paper uses these models to cater to the needs of the problem presented in the shared tasks.

Many transformer models were trained using the training data and dev data to test the performance, as shown in Table 2. After testing the performance of various transformer models, the top three models with the best performance were selected. The models, XLM RoBERTa base (XLMR) (Conneau et al., 2019), multilingual-cased BERT base (mBERT) (Devlin et al., 2019), MuRIL cased (MuRIL) (Khanuja et al., 2021) were selected. The selected models were trained after concatenating the train and the dev dataset for the final predictions.

XLMR is an unsupervised model which has been trained on 100 different languages. This model was based on Facebook's 2019 RoBERTa (Liu et al., 2019) model. This is a large multi-lingual model which was trained on 2.5TB of filtered data from CommonCrawl.

MuRIL cased temp model is an NLP model that has been trained in the transformers library implemented using Python.

mBERT is a self-supervised transformer model which was pre-trained on a large multilingual corpus. The model was not trained on data labeled by humans instead, it was trained on raw texts. This model was trained on 104 languages with the largest Wikipedia. This model is case sensitive in nature.

It was observed that the performance of the selected models used in the ensemble model suffered after pre-processing the text, where the pre-processing included the removal of numbers, special characters and emojis. Hence, no pre-processing was done before training the models. The performances of the various models that were implemented have been shown in Table 2 both before and after preprocessing the training data and testing its performance on the dev data. The Adam optimizer was used with a learning rate of 1e-5 and cross entropy function was used as the loss function for all the models.

For performing tokenization, different tokenizers were used, which were specific to each model. The XLMRoBERTaTokenizer was used for XLMR, the BertTokenizer was used for mBERT, and the AutoTokenizer was used for MuRIL.

For evaluating the label, as shown in Figure 1, the data was first passed through the three models individually. Then the predictions of these three models were combined by using majority voting. The label with the highest frequency was finally predicted as the output of the ensemble model.

5 Results and Discussion

The ensemble model was designed by selecting the top three models with the best performance on the training data. The performance of all the models that were implemented on the training data has been shown in Table 2.

The base models and transformer models were both trained and tested. The transformer models consistently performed better than the base machine learning models. Each of the transformer models was trained for 5 to 50 epochs each. The highest F1 score and its corresponding accuracy have been mentioned in Table 2 for the transformer models.

The proposed ensembling technique achieved the highest F1 score of 0.82. This also shows that combining the various transformer-based techniques can lead to improved performance.

6 Conclusion and Future Work

In this work, an ensembling technique was proposed to automatically detect whether a given text in Tamil contains caste and migration hate speech for the Caste and Migration Hate Speech Detection in Tamil shared task by the LT-EDI 2024 workshop at EACL 2024. The technique was based on transformer models and utilized transfer learning.

The performance of the ensemble model can be further improved by taking the predictions from more transformer models or other traditional machine learning and deep learning techniques. Also, taking a weighted vote of the models according to their performance on the training data can help give better results than majority voting, where each model is given equal importance irrespective of their performance relative to the other models.

References

Snehaan Bhawal, Pradeep Roy, and Abhinav Kumar. 2021. Hate speech and offensive language identification on multilingual code mixed text using BERT. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR.*

Asoka Chakravarthi and Bharathi Raja. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi. 2020. "HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion". In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. 2022. "Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion". In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. "IIITK@DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada". In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

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.

Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, B Bharathi, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya Sambath Kumar, Thomas Mandl, Rahul Ponnusamy, Vasanth Palanikumar, et al. 2022. Overview of the Shared Task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages.

Kogilavani Shanmugavadivel, Malliga Subramanian, Shri Durga R, Srigha S, Sree Harene J S, and Yasvanth Bala P. 2023. "KEC_AI_NLP@DravidianLangTech: Abusive Comment Detection in Tamil Language". In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. *Computer Speech Language*, 76:101404.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.