MEnTr@LT-EDI-2024: Multilingual Ensemble of Transformer Models for Homophobia/Transphobia Detection

Adwita Arora[1], Aaryan Mattoo[1], Divya Chaudhary[2], Ian Gorton[2] and **Bijendra Kumar[1]**
[1] Netaji Subhas University of Technology, New Delhi, India
[2] Northeastern University, Boston, MA 02115, United States
{adwita.ug20, aaryan.mattoo.ug20}@nsut.ac.in
{d.chaudhary, i.gorton}@northeastern.edu
bizender@nsut.ac.in

Abstract

Detection of Homophobia and Transphobia in social media comments serves as an important step in the overall development of Equality, Diversity and Inclusion (EDI). In this research, we describe the system we formulated while participating in the shared task of Homophobia/Transphobia detection as a part of the Fourth Workshop On Language Technology For Equality, Diversity, Inclusion (LT-EDI-2024) at EACL 2024[1]. We used an ensemble of three state-of-the-art multilingual transformer models, namely Multilingual BERT (mBERT), Multilingual Representations for Indic Languages (MuRIL) and XLM-RoBERTa to detect the presence of Homophobia or Transphobia in YouTube comments. The task comprised of datasets in ten languages - Hindi, English, Telugu, Tamil, Malayalam, Kannada, Gujarati, Marathi, Spanish and Tulu. Our system achieved rank 1 for the Spanish and Tulu tasks, 2 for Telugu, 3 for Marathi and Gujarati, 4 for Tamil, 5 for Hindi and Kannada, 6 for English and 8 for Malayalam. These results speak for the efficacy of our ensemble model as well as the data augmentation strategy we adopted for the detection of anti-LGBT+ language in social media data.

1 Introduction

Homophobia is defined as intentional discrimination against those who identify as a part of the LGBT+ community. It can be demonstrated in many ways, which can include abuse or social ignorance. Transphobia, on the other hand, refers to the targeted hatred towards transgender individuals whose current gender identity and the one assigned to them during birth differ. Both of these forms of hate speech have negative repercussions on the mental health as well as the overall well-being of people who are a part of the LGBT+ community (Chakravarthi et al., 2022a). This highlights a critical need to build systems that identify this form of prejudice and bigotry.

Pre-trained language models (PLMs) like BERT (Devlin et al., 2018) and GPT (Brown et al., 2020), built on transformer architectures have gained recognition for their ability to interpret languages in a manner similar to humans by displaying state-of-the-art results in many NLP tasks such as document classification and language modelling. PLMs undergo unsupervised training on a large corpus of text data which can then be fine-tuned on domain and task-specific corpora for downstream tasks, such as the shared task on Homophobia and Transphobia detection by LT-EDI@EACL 2024. BERT, specifically, introduced the concept of bidirectional context understanding which considers both the succeeding and preceding word for a particular word to capture a more elaborate and nuanced meaning within the language. For our system, we propose an ensemble consisting of three such popular BERT-based transformer architectures, namely Multilingual BERT (mBERT) (Devlin et al., 2018), Multilingual Representations for Indic Languages (MuRIL) (Khanuja et al., 2021) and XLM-RoBERTa (Conneau et al., 2019).

1.1 Task Description

As specified in (Chakravarthi et al., 2024), participants of this shared task were required to submit systems that classify a given YouTube comment into one of the three categories - Homophobia, Transphobia or None. We were provided with the train and development datasets containing manually annotated posts in English, Hindi, Malayalam, Tamil, Telugu, Kannada, Marathi, Gujarati and Spanish. The dataset described by Kumaresan et al. (2023) forms the seed data for this task. This year, the workshop also introduced a code-mixed dataset on Tulu. Being an under-resourced language, Tulu lacks extensive data and resources

[1] https://sites.google.com/view/lt-edi-2024/

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 259–264

Language		Non anti-LGBT+ content	Homophobia	Transphobia
Tamil	train	2,064	453	145
	dev	507	118	41
	test	634	152	47
Telugu	train	3,496	2,907	2,647
	dev	747	588	605
	test	744	624	571
Kannada	train	4,463	2,765	2,835
	dev	955	585	617
	test	951	599	606
Gujarati	train	3,848	2,267	2,004
	dev	788	498	454
	test	794	510	436
Spanish	train	700	250	250
	dev	200	93	93
	test	300	150	150
Hindi	train	2,423	45	92
	dev	305	2	13
	test	308	3	10
English	train	2,978	179	7
	dev	748	42	2
	test	931	55	4
Malayalam	train	2,468	476	170
	dev	937	197	79
	test	674	140	52
Marathi	train	2,572	551	377
	dev	541	129	80
	test	569	112	69

Table 1: Statistics of the train, dev and train dataset

		Non H/T Content	H/T Content
Tulu	train	542	188
	test	312	67

Table 2: Statistics of the Tulu train and test dataset

for language models. This scarcity leads to a few-shot learning scenario. For the Tulu task, we were required to build a binary classifier that predicts whether a post contains hate-speech relating to homophobia or transphobia. The overall task hence is to develop a multiclass (binary in case of Tulu) classifier that predicts whether a given post contains instances of homophobia or transphobia in 10 different language categories. The systems were weighed using the average macro F1 score for each language across all classes on the test dataset.

2 Related Work

Transformer models have been popular in various classification tasks, including hate speech detection. Roy et al. (2021) experimented with the XLM-RoBERTa model for hate-speech detection in Twitter data in English, German and Hindi. Top submissions to competitions like HASOC (Hate Speech and Offensive Content Identification in Multiple Languages) which provide datasets for hate-speech detection in a multilingual setting also utilised transformer models, such as Farooqi et al. (2021) who used IndicBERT, XLM-RoBERTa and Multilingual BERT with hard voting.

The task presented at this workshop is the third shared task on Homophobia and Transphobia detection in social media comments. In the previous shared tasks, Chakravarthi et al. (2022b) and Chakravarthi et al. (2023), the majority of the submissions received used transformer models, such as Nozza (2022) who used weighted majority voting on the predictions received from BERT, RoBERTa and HateBERT and Maimaitituoheti (2022) who used the pre-trained transformer model RoBERTa for classification. Other submissions also experimented with neural networks and support vector machines such as (García-Díaz et al., 2022) and (Ashraf et al., 2022) respectively. Bhandari and Goyal (2022) experimented with various multilingual BERT models, including mBERT, XLM-RoBERTa, IndicBERT and HateBERT, with a data augmentation strategy of random insertion, deletion or swapping of words in a sentence.

3 Methodology

Language	Homophobia	Transphobia
Tamil	1,146	1,049
Telugu	2,907	2,647
Kannada	2,765	2,835
Gujarati	2,267	2,004
Spanish	316	316
Hindi	837	820
English	1,223	953
Malayalam	1,277	1,189
Marathi	1,209	1,259

Table 3: Train corpora after augmentation

The distribution of labels in the train and dev splits are shown in Table 1 and Table 2. From looking at the balance of classes in the train dataset, it is inferred that the Homophobia and Transphobia classes are highly imbalanced, especially for Hindi, English, Tamil, Malayalam, Marathi and Spanish.

3.1 Handling Class Imbalance

To provide a balance between the classes of the dataset across all languages, we use a translation strategy where we take the positive samples from Kannada, Gujarati and Telugu and translate them into each of our target languages i.e., Hindi, English, Tamil, Malayalam, Marathi and Spanish. This is done for both Homophobia and Transphobia classes. We used the `googletrans`[2] library in Python for the translation process. The distribution of the modified dataset is shown in Table 3.

3.2 Ensemble of Transformer Models

For this classification task, we propose an ensemble of three of the most popular multilingual transformer models built on top of the BERT architecture, as described below:

- **mBERT:** Multilingual BERT (mBERT) (Devlin et al., 2018) is a pre-trained model which is trained using data belonging to 104 languages. We used the `bert-base-multilingual-case`[3] pre-trained model.

- **MuRIL:** Multilingual Representations for Indian Languages (MuRIL) (Khanuja et al., 2021) is a BERT-based model that has been

[2]https://pypi.org/project/googletrans/

[3]https://huggingface.co/bert-base-multilingual-cased

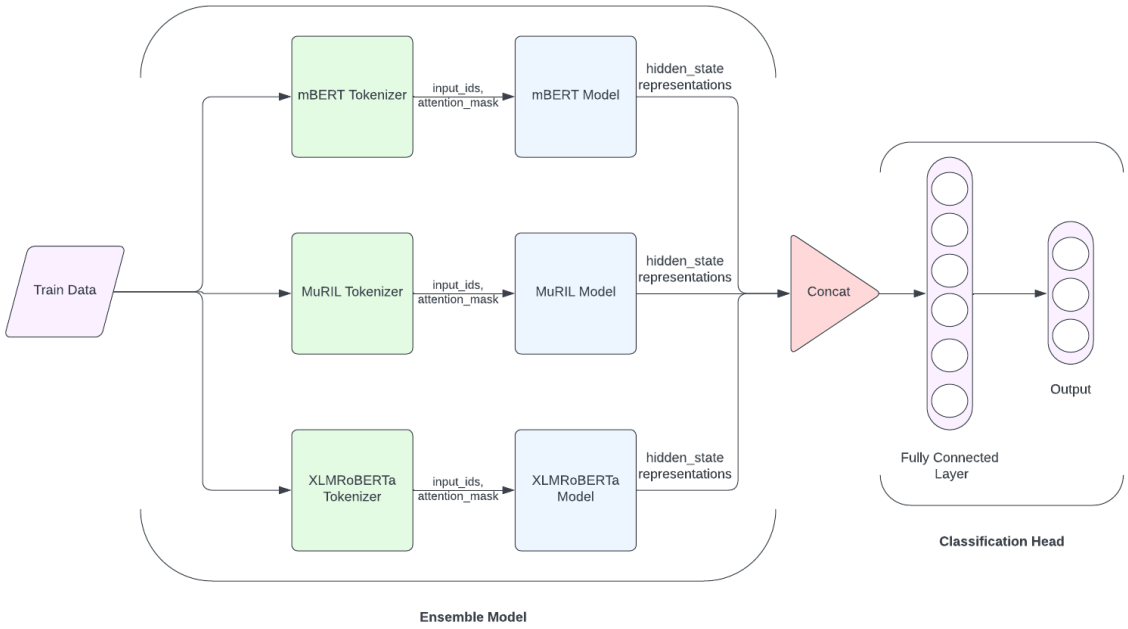

Figure 1: System Architecture

pre-trained on 16 Indian languages. We used the `google/muril-base-cased`[4] model. We removed the MuRIL layer while fine-tuning for the Spanish language condition, given the fact that MuRIL is pre-trained on Indic Languages specifically.

- **XLMRoBERTa:** XLMRoBERTa (Conneau et al., 2019) is a cross-linguistic pre-trained linguistic model built by Meta. We used the `xlm-roberta-base`[5] model.

These models are used with the help of the HuggingFace library [6] for transformer models.

Figure 1 depicts our system, where the train dataset is first tokenized and fed to its corresponding transformer model. The hidden state representation obtained from each of the three models is concatenated and fed as input to a simple classification head consisting of a feed-forward neural network which outputs the predicted class.

We fine-tune the ensemble model on the Google Colab GPU on the train dataset for each language task. We train each language model for 3 epochs using Binary Cross Entropy as the loss function and AdamW (Loshchilov and Hutter, 2017) as the optimizer. We kept the learning rate at 2e-5. The fine-tuned model then generated the predicted classes for test data in each language, which was submitted for evaluation.

4 Results and Discussions

The results obtained for each language task are given in Table 4. This shows the final average macro F1 score obtained for each language on the test dataset as shared by the organizers. The best results were seen in the case of Spanish and Tulu where we achieved a rank of 1. In the Telugu language task, our system ranked second with an average macro F1 score of 0.960 which also was the overall best average macro F1 score across all languages for our system. This can be explained by Telugu having the best class distribution across all languages. Even though we used translation as a data augmentation strategy, it does not ensure that all the linguistic features of the source text are retained in the target text. Our system performs well with Marathi and Gujarati as well, ranking third with average macro F1 scores of 0.488 and 0.960 respectively. For the rest of the languages we see varying performance with Tamil ranking fourth with an average macro F1 score of 0.746

[4]https://huggingface.co/google/muril-base-cased
[5]https://huggingface.co/xlm-roberta-base
[6]https://huggingface.co/

Language	Average Macro F1	Rank
Tulu	0.707	1
Spanish	0.582	1
Telugu	0.960	2
Gujarati	0.960	3
Marathi	0.488	3
Tamil	0.746	4
Hindi	0.325	5
Kannada	0.935	5
English	0.407	6
Malayalam	0.744	8

Table 4: Results showing the average macro F1 score

and Hindi and Kannada ranking fifth with an average macro F1 score of 0.325 and 0.935. For English and Malayalam the performance was not at par with the other languages with ranks 6 and 8 and average macro F1 scores 0.407 and 0.744 respectively. There is a direct link between the average macro F1 score and the distribution of classes in the train dataset, even after data augmentation. The translation schemes, while improving the diversity of the dataset to a certain extent, do not guarantee an improvement in the quality of the dataset. Languages like Hindi and English that had the poorest class balance in the train dataset also resulted in the poorest average macro F1 scores of 0.325 and 0.407 on the test dataset. For the languages having a more diverse distribution like Telugu and Gujarati, we also see a higher average macro F1 score. Results for Tulu are also impressive considering that none of the pre-trained BERT models were trained on corpora containing text data in Tulu. However, given the linguistic and phonetic similarities between Tulu, Kannada and Malayalam, the ensemble model was able to capture the features of this language to a certain extent, resulting in an average macro F1 score of 0.707, ranking first.

5 Conclusions and Future Work

Our submission for the shared task on Homophobia and Transphobia detection in social media comments demonstrates how pre-trained language models (PLMs) specifically those built on a BERT-based architecture can be effectively used in the case of text classification. Our ensemble model consisting of mBERT, MuRIL and XLMRoBERTa, has shown consistent results by achieving the top three ranks for 5 language tasks, ranking first for Spanish and the under-resourced language Tulu. We have been able to achieve average macro F1 scores of 0.707, 0.582, 0.960, 0.960, 0.488, 0.746, 0.325, 0.935, 0.407 and 0.744 for Tulu, Spanish, Telugu, Gujarati, Marathi, Tamil, Hindi, Kannada, English and Malayalam respectively. In the future, we would like to experiment with the following aspects in further detail :

- **Better data augmentation strategies:** Simple translation from one language to another does not consider the linguistic nuances of these languages, which is required to build a diverse and high-quality dataset. We would like to experiment with more sophisticated, language-dependent data augmentation strategies.
- **Attention mechanisms:** Addition of attention modules to the ensemble model to further capture complex positional dependencies in multilingual code-mixed data.

6 Limitations

The data presented in the shared task comprised 10 different languages, each with its own linguistic and cultural nuance, and we recognise that bringing forth a common end-to-end approach for text classification may miss some of these nuances. However, the system we presented stands a baseline which can easily be extended to include language and context specific modules before training.

References

Nsrin Ashraf, Mohamed Taha, Ahmed Abd Elfattah, and Hamada Nayel. 2022. NAYEL @LT-EDI-ACL2022: Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 287–290, Dublin, Ireland. Association for Computational Linguistics.

Vitthal Bhandari and Poonam Goyal. 2022. bitsa_nlp@LT-EDI-ACL2022: Leveraging Pre-trained Language Models for Detecting Homophobia and Transphobia in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 149–154, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Saranya Rajiakodi, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Kishore Kumar Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, S Malliga, Paul Buitelaar, Salud Mar ıa Jimenez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Nitesh Jindal, et al. 2023. Overview of second shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 38–46.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of The Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

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. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

José García-Díaz, Camilo Caparros-Laiz, and Rafael Valencia-García. 2022. UMUTeam@LT-EDI-ACL2022: Detecting homophobic and transphobic comments in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 140–144, Dublin, Ireland. Association for Computational Linguistics.

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.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, page 100041.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Abulimiti Maimaitituoheti. 2022. ABLIMET @LT-EDI-ACL2022: A RoBERTa based Approach for Homophobia/Transphobia Detection in Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 155–160, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble Modeling for Homophobia and Transphobia Detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. *arXiv preprint arXiv:2101.03207*.