Sammaan@LT-EDI-ACL2022: Ensembled Transformers Against Homophobia and Transphobia

Ishan Sanjeev Upadhyay and **KV Aditya Srivatsa** and **Radhika Mamidi**
International Institute of Information Technology, Hyderabad
{ishan.sanjeev, k.v.aditya}@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

Hateful and offensive content on social media platforms can have negative effects on users and can make online communities more hostile towards certain people and hamper equality, diversity and inclusion. In this paper, we describe our approach to classify homophobia and transphobia in social media comments. We used an ensemble of transformer based models to build our classifier. Our model ranked 2nd for English, 8th for Tamil and 10th for Tamil-English.

1 Introduction

Social media platforms allow people from all walks of life to connect with each other. However, abusive and hateful content on these platforms can take a psychological toll on its users (Wypych and Bilewicz, 2022) (Tynes et al., 2008). Lesbian, gay, bisexual and transgender individuals are more vulnerable to mental illness as compared to their heterosexual peers (Gilman et al., 2001) (Marshal et al., 2011) (Reisner et al., 2015). Hence, it becomes even more important to be able to detect such hateful content for vulnerable individuals.

There has been a lot of work done in the domain of hate speech detection (Malmasi and Zampieri, 2017) (Burnap and Williams, 2016). There has also been work on hate speech intervention (Qian et al., 2019). Shared tasks like SemEval 2019 Task 6 have focused on identifying and categorizing offensive language on social media (Zampieri et al., 2019). Datasets for this task have been created in multiple languages as well. Bohra et al. (2018) created a Hindi-English code mixed text dataset for hate speech detection from tweets on Twitter. Mubarak et al. (2021) created a 1000 tweets Arabic dataset for offensive language detection with special tags for vulgarity and hate speech. Sigurbergsson and Derczynski (2020) created a Danish hate speech detection dataset containing 3600 user generated comments social media websites. There have been datasets created for Greek (Pitenis et al., 2020) and Turkish (Çöltekin, 2020) as well. Chakravarthi et al. (2021a) created a code-mixed Tamil,Malayalam and Kannada dataset for offensive language identification. Support vector machines, long short-term memory networks, convolutional neural networks and now transformer based architectures have been used to detect hate speech. However, there has not been much work in trying to specifically identify homophobic or transphobic text. In this paper, we will describe our approach for classifying transphobic and homophobic comments in the dataset provided by Chakravarthi et al. (2021b) as a part of the shared task on homophobia and transphobia detection in social media comments Chakravarthi et al. (2022).

2 Dataset Description

The dataset consists of a total of 15,141 comments in 3 languages: English, Tamil and Tamil-English code-mixed (refer to Table 1 for data distribution). Each comment has one of three labels "Homophobic", "Transphobic" and "Non-anti-LGBT+ content" (label distribution in Table 2).

3 Methodology

In this section we will describe the models used in our experimentation.

- **BERT**: BERT (Devlin et al., 2019) is a Transformer-based language model. It consists of layered encoder units, each with a self-attention layer followed by fully-connected layers. It is trained using the Masked Language Modelling (MLM) task as well as the Next Sentence Prediction (NSP) task. For this shared task, we have used the pretrained bert-base-uncased model from HuggingFace (Wolf et al., 2019).

- **RoBERTa**: RoBERTa (Liu et al., 2019) is a Transformer-based language model which

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 270 - 275

Language	Number of comments	Number of tokens	Number of characters
English	4,946	82,111	438,980
Tamil	4,161	197,237	539,559
Tamil-English	6,034	66,731	435,890
Total	15,141	346,079	1,414,429

Table 1: Distribution of comments in English, Tamil and Tamil-English.

Class	English	Tamil	Tamil English
Homophobic	276	723	465
Transphobic	13	233	184
Non-anti-LGBT+ content	4,657	3,205	5,385
Total	4,946	4,161	6,034

Table 2: Distribution between Homophobic, Transphobic and Non-anti-LGBT+ content.

improves upon the BERT architecture along several metrics offered by the GLUE benchmark (Wang et al., 2019). It is not trained on the NSP task and involves dynamic masking for the MLM task. It is also trained over a much larger dataset with longer sentence lengths. For this shared task, we have used the pretrained roberta-base model.

- **HateBERT** (Caselli et al., 2021) is a retrained BERT model to detect abusive language in English. It is trained on large amounts of banned Reddit comments extracted from the RAL-E dataset. It has been shown to outperform the BERT model in several hate-speech detection tasks.

- **IndicBERT**: IndicBERT (Kakwani et al., 2020) is an ALBERT Transformer encoder (Lan et al., 2020) finetuned on data from 12 major Indian languages, including 549M tokens of Tamil. Despite having significantly lower parameters than other multilingual encoders such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020), it outperforms them on several metrics of the IndicGLUE benchmark (Kakwani et al., 2020). We have used the IndicBERT model as a TLM for the Tamil and Tamil-English tracks.

- **XGBoost Random Forest Classifier**: Random Forest Classifiers (Ho, 1995) are meta estimators which consist of numerous decision trees, each fit upon a subset of features from a subset of rows of the data. The ensemble of many such weak learners tends to outperform a single large decision tree. The low correlation between the constituent trees also provides for more feature coverage and curbs over-fitting. For this shared task, we use XGBoost's implementation of Random Forest Classifiers (Chen and Guestrin, 2016).

- **Bayesian Optimization**: The aim of any hyperparameter optimization strategy is to find the hyperparameter set which fetches the best value over the object function. Bayesian Optimization (Mockus, 1989) is an iterative optimization algorithm that aims to minimize the number of hyperparameter sets that must be evaluated before arriving at the optimal distribution. It has been shown to generate optimal solutions in significantly fewer iterations than traditional methods such as grid search. For this task, we have used the Python library: bayesian-optimization (Fernando, 2014).

4 Experiments and Results

The only pre-procesesing step done on the dataset before training was the change of emojis to text using the demoji library in python [1]. Our pipeline comprises an ensemble of several Transformer-based language models (TLM), namely: BERT, RoBERTa, and HateBERT for the English track and IndicBERT for the Tamil and Tamil-English tracks. Three copies of each TLM are used with different parameter initializations in each track. This allows for the copies to capture different features of the data. In addition to this, for each track, a layer of attention is applied to each constituent encoder layer outputs of the TLMs. This is necessary since

[1] https://pypi.org/project/demoji/

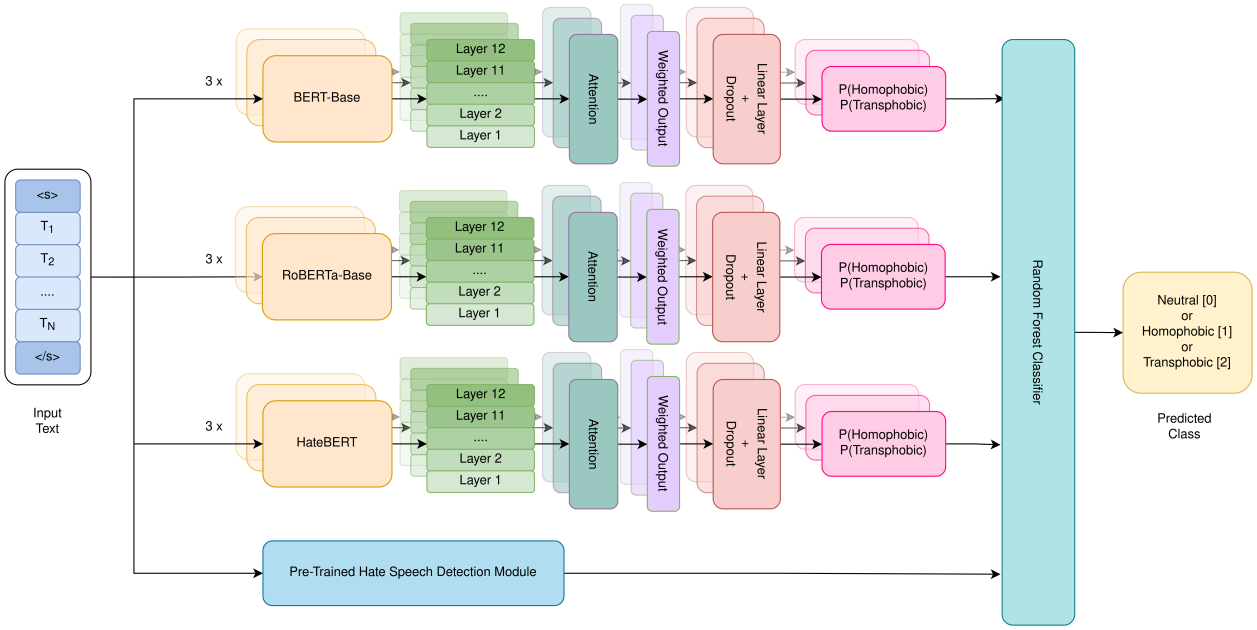

Figure 1: Schematic overview of the architecture of our model.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1
BERT	0.92	0.48	0.42	0.44	0.9	0.92	0.91
RoBERTa	0.93	0.64	0.36	0.36	0.93	0.94	0.9
HateBERT	0.94	0.56	0.43	0.47	0.92	0.94	0.92
Ensemble	**0.94**	**0.52**	**0.47**	**0.49**	**0.93**	**0.94**	**0.94**

Table 3: Classification results of various models used on the English dataset.

each layer captures a different kind of information, which are variably relevant for our task. The weighted and combined output from the attention layer is then passed through a final linear layer and dropout layer ($p = 0.3$), followed by a Softmax operation to generate the predicted probabilities of detecting homophobic content in the given input text.

In the English track, we also use a pretrained hate-speech detection model implemented on HuggingFace (Wolf et al., 2019). Architecturally, is a ByT5-Base model (Xue et al., 2021) finetuned on HuggingFace's tweets_hate_speech_detection dataset (Sharma, 2019).

The prediction probabilities are generated by each model of a track are passed as input features to a Random Forest Classifier. This helps further optimize our predictions by weighing the importance of the different architectures for the task.

Each of the TLM pipelines was finetuned upon Cross Entropy loss using AdamW optimizer (Loshchilov and Hutter, 2017) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with an initial learning rate of $2e^{-5}$ for 6 epochs each using a linear scheduler. The epoch checkpoint with the highest validation F1 score was selected for further use. The hyperparameters of the Random Forest Classifier were estimated using 10 seeds and 100 iterations of Bayesian Optimization. The ensemble classifier was trained with a learning rate of 1.0.

As can be seen in Table 3, our ensemble model performed better than the individually trained models giving a macro F1 score of 0.49 which was the 2nd highest macro F1 score in the shared task. This model also had the highest weighted F1 score in the task. The IndicBERT ensembles trained on the Tamil and Tamil-English dataset give us a macro F1 score of 0.55 and 0.35 and a weighted F1 score of 0.86 and 0.83 respectively (refer Table 4). The Tamil and Tamil-English model ranked 8th and 10th respectively.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1
Tamil-English	0.83	0.34	0.35	0.35	0.82	0.83	0.83
Tamil	0.88	0.52	0.58	0.55	0.85	0.88	0.86

Table 4: Classification results of IndicBERT finetuned on the Tamil-English and Tamil dataset.

5 Conclusion and Future Work

In this paper, we described our approach for homophobia and transphobia detection in English, Tamil and Tamil-English. We used an ensemble of three transformed based models along with a pre-trained hate detection model to do the classification for English. Our model was ranked 2nd for the English classification task. For the Tamil and Tamil-English dataset three copies of the IndicBERT model was used to make our ensemble based model. The models placed 8th and 10th for Tamil and Tamil-English model respectively.

In the future, we can use data augmentation methods like paraphrasing and back translation to increase the diversity and quantity of homophobic and transphobic text. We can also incorporate transliteration into the pipeline for Tamil-English code mixed text since IndicBERT is not trained on code mixed text. We could also try to finetune transformers pre-trained on code mixed data.

References

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1).

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021a. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021b. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Nogueira Fernando. 2014. Bayesian Optimization: Open source constrained global optimization tool for Python.

S E Gilman, S D Cochran, V M Mays, M Hughes, D Ostrow, and R C Kessler. 2001. Risk of psychiatric disorders among individuals reporting same-sex sexual partners in the national comorbidity survey. *American Journal of Public Health*, 91(6):933–939.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

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.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Michael P. Marshal, Laura J. Dietz, Mark S. Friedman, Ron Stall, Helen A. Smith, James McGinley, Brian C. Thoma, Pamela J. Murray, Anthony R. D'Augelli, and David A. Brent. 2011. Suicidality and depression disparities between sexual minority and heterosexual youth: A meta-analytic review. *Journal of Adolescent Health*, 49(2):115–123.

Jonas Mockus. 1989. *Bayesian approach to global optimization*. Kluwer Academic.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Sari L Reisner, Ralph Vetters, M Leclerc, Shayne Zaslow, Sarah Wolfrum, Daniel Shumer, and Matthew J Mimiaga. 2015. Mental health of transgender youth in care at an adolescent urban community health center: a matched retrospective cohort study. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 56(3):274–9.

Roshan Sharma. 2019. $\text{tweets}_{h}ate_{s}peech_{d}etectiondatasetsathuggingface$.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Brendesha M. Tynes, Michael T. Giang, David R. Williams, and Geneene N. Thompson. 2008. Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health*, 43(6):565–569.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Michał Wypych and Michał Bilewicz. 2022. Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among ukrainian immigrants in poland. *Cultural Diversity and Ethnic Minority Psychology*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *CoRR*, abs/2105.13626.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.