ABLIMET @LT-EDI-ACL2022: A RoBERTa based Approach for Homophobia/Transphobia Detection in Social Media

Abulimiti Maimaitituoheti, Yang Yong, Fan Xiaochao
Xinjiang Normal University, China
{1149654712, 68523593, 37769630}@qq.com

Abstract

This paper describes our system that participated in LT-EDI-ACL2022-Homophobia/Transphobia Detection in Social Media. Sexual minorities face a lot of unfair treatment and discrimination in our world. This creates enormous stress and many psychological problems for sexual minorities. There is a lot of hate speech on the internet, and homophobia/transphobia is one against sexual minorities. Identifying and processing homophobia/transphobia through natural language processing technology can improve the efficiency of processing it, and can quickly screen out it on the Internet. The organizer of the competition constructs a homophobia/ transphobia detection dataset based on YouTube comments for English and Tamil. We use a RoBERTa-based approach to conduct our experiments on the dataset of the competition, and get better results.

1 Introduction

At present, the Internet is full of various hate speeches, including racial discrimination, religious hostility, mutual hostility between political groups, and discrimination against sexual minorities. The discrimination against sexual minorities is called homophobia/transphobia. Homophobia and transphobia are two concepts that are similar and different. Homophobia refers to unwarranted fear, hatred, and unfair treatment of homosexuals, and transphobia refers to disgust and discrimination against transgender people. Homophobia/ transphobia will bring serious psychological stress to LGBTQ people, making them unable to participate in social activities normally, and even causing them serious mental illness. Therefore, the quick and efficient identification and screening of homophobia/transphobia on the Internet will help to clean up cyberspace, build a healthy and harmonious Internet community, and help more people realize the unfair treatment of LGBTQ groups.

Shared Task on Homophobia/Transphobia Detection in social media comments at LT-EDI 2022- ACL 2022 is a classification task. The organizer of the competition constructed a homophobia/transphobia detection dataset in English, Tamil, and Tamil-English based on YouTube comments. The model needs to determine whether the target data contains homophobia/ transphobia. And if so, which type of homophobia/ transphobia is included. We used RoBERTa (Liu et al.,2019) as our pre-trained language model and fine-tuned it for the task. In our experiment, we process the target data by the pre-trained language model, the output was normalized firstly by a layer normalization module， then we use two fully-connected layers and between them there is a layer normalization operation. We use the cross-entropy as our loss function, and optimize it by AdamW (Loshchilov and Hutter, 2019). Through the above steps, we complement the identification and classification of homophobia/transphobia in the target data.

We participated in all of the English, Tamil, and Tamil-English homophobia/transphobia detection subtasks, and we use a version of RoBERTa pre-trained on the corresponding linguistic data for each language. By training the model on train data and validating it on development data, we achieved better results on test data. Specifically, we achieved a 0.57 macro f1-score on the English

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 155 - 160

Figure 1: Architecture of homophobia/transphobia detection model

subtask and ranked 1st among all participating teams, achieved a 0.75 macro f1-score on the Tamil subtask and ranked 5th among all participating teams, achieved a 0.53 macro f1 score on the Tamil-English subtask and ranked 6th among all participating teams.

2 Background

In this section, we introduce the relevant background of the Shared Task on Homophobia/ Transphobia Detection in social media comments at LT-EDI 2022- ACL 2022, including the details of the task and the related research on homophobia/transphobia detection.

2.1 Problem Description

Shared Task on Homophobia/Transphobia Detection in social media comments at LT-EDI 2022- ACL 2022 is a classification task in English, Tamil, and Tamil-English. The organizer of the competition constructed the homophobia/ transphobia detection dataset based on the homophobia and transphobia identification dataset (Chakravarthi et al.,2019). The target data is a YouTube comment which may contain one or more homophobia/transphobia. A model needs to determine whether the comment contains homophobic or transphobic information and classify the comment into one of the 3 labels: Homophobic, Transphobic, or Non-anti-LGBT+ content. The Homophobic label refers to the comment containing homophobic information, the Transphobic label then refers to the comment containing transphobic information and the Non-anti-LGBT+ content label refers to that the comment doesn't contain homophobic or transphobic information. For example:

- They harass everyone on the bus and do this for living. -Homophobic
- Hey seriously I thought She was Transgender. -Transphobic
- Don't worry everything will be solved soon. - Non-anti-LGBT+ content

2.2 Related Works

So far, people have carried out a lot of research on emotion recognition, hate speech detection in low resource and code-mixed data, researched homophobia/transphobia from different perspectives such as linguistics, psychology, sociology, and pedagogy, and clarified the harm and trouble that homophobia/transphobia brings to sexual minorities. Divyansh (2021) collected Hindi-English code-mixed twitters and comments from Twitter and video streaming platforms by using data scraping tools, constructed an emotion recognition dataset by manually annotating all twitters, and comments, and conducted emotion recognition experiments by using models SVM, LSTM, etc. Ravindra and Raviraj (2021) conducted hate speech detection experiments on a code-mixed twitter dataset by using Multilingual BERT (Telmo et al., 2019) and Indic-BERT (Divyanshu et al., 2020) and achieved better results. Fernando et al (2020) clarified the distress and harm that the sexual minorities suffered and propose alternatives to providing better and more equitable education for sexual minorities. Gamez and Daniel (2021) examine discrimination and prejudice against sexual minorities on the Internet and discuss the mental health of LGBTQ adolescents. Lin et al (2021) made a systematic survey on The mental health of transgender and gender non-conforming people in China. However, from the perspective of computer linguistics, there are few papers on the identification and screening of homophobia and transphobia. Chakravarthi et al (2019) constructed a multilingual homophobia/ transphobia detection dataset based on YouTube comments and made homophobia/transphobia detection experiments by a lot of models like SVM, LSTM, BERT (Devlin et al.,2019), etc.

3 System Overview

In this section, we will introduce our approach to the task, the multi-label homophobia/transphobia detection task, which we solve using a fine-tuning approach of the pre-trained language model. Specifically, we process the target data by the pre-trained language model, the output was normalized firstly by a layer normalization module , then we use two fully-connected layers and between them with a layer normalization operation.

The model architecture is shown in Figure 1. Input is the target data to be processed, and the

Table 1: Statistical Details of the Data sets

Language	Data Set	Shorter Than 128 Words	Between 128 And 192 Words	Longer Than 192 Words
English	Train Set	3126	25	3
	Development Set	776	4	3
Tamil	Train Set	2196	222	236
	Development Set	548	48	59
Tamil-English	Train Set	3785	49	22
	Development Set	945	9	2

pre-trained language model (PLM) processes the input data and outputs the result to the layer normalization module. To prevent internal covariate shift, we follow the pre-trained language model and first fully connected layer with a layer normalization module. The pre-trained language model we use is RoBERTa-base for English subtask, Tamil-RoBERTa for Tamil, and Tamil-English subtasks. The output tensor size of the pre-trained language model is 768, while our target label is only 3, the difference between the two is large, so we connect two full connection layers behind the pre-trained language model layer, where the output tensor size of the first fully
connected layer is 64, and the output tensor size of the second fully connected layer is 3,the number of target labels.

For the loss function, we use the cross-entropy provided by the PyTorch (Paszke et al.,2019) framework, and use Adamw (Loshchilov and Hutter, 2019) as optimizer, which is an improved version of the Adam (Kingma and Ba, 2017) optimizer. Due to the huge parameters of the pre-trained language model, it is easy to overfit when using the Adam optimizer, while Adamw uses L2 regularization to reduce overfitting, which can significantly improve the generalization ability of the model.

4 Experimental Setup

In this section, we will introduce the relevant design, parameter settings, and experimental environment of the experiments. In terms of hardware, we use a laptop with a GTX 1650 graphics card for model training. On the software side, we use PyTorch (Paszke et al.,2019) and HuggingFace (Wolf et al., 2020) library to code the tasks.

4.1 Statistical Analysis

Shared Task on Homophobia/Transphobia Detection in social media comments at LT-EDI 2022- ACL 2022 provided datasets for English, Tamil, and Tamil-English, and these datasets was splitted into train, development, and test set. Train and development data was provided with labels and test set only provided the target text. To set reasonable parameters for our experiment, we made a statistical analysis for the train and development sets of the data sets, the details are shown in table1.

Combined with the data statistics and the experimental hardware environment, considering the experimental performance and cost, we set the max data length processed by the pre-trained language model for English and Tamil-English to 128, and set for Tamil to 192. It means that the redundant part of a target data which is longer than 128 for English and Tamil-English, 192 for Tamil will be discarded and will not participate in model training and testing.

Table 2: Details of the train datasets before and after balanced

Language	Befor Balanced				After Balanced			
	Homo	Trans	LGBT	Total	Homo	Trans	LGBT	Total
English Train	157	6	3001	3164	3001	3001	3001	9003
	485	155	2022	2662	2022	2022	2022	6066
Tamil Train	311	112	3438	3861	3438	3438	3438	10314
	157	6	3001	3164	3001	3001	3001	9003
Tamil-English Train	485	155	2022	2662	2022	2022	2022	6066
	311	112	3438	3861	3438	3438	3438	10314

4.2 Data Balancing

The datasets provided by the organizer of the competition are unbalanced datasets. If we use the unbalanced dataset to train our model directly, the model will only learn features from a larger number of categories and will ignore features of a smaller number of categories, so it is necessary to balance the datasets. So we use the class RandomOverSampler from the imbalanced-learn (Guillaume et al.,2017) library to balance the train datasets. The RandomOverSampler class will simply copy-paste the data for a smaller number of categories so that each category in the dataset has an equal amount of data. Although this method is simple, it is more effective for model training. We only balanced the train datasets for the three languages, and use the original development and test datasets to validate and test our model. The details of the train datasets before and after balancing them are shown in table 2.

In table 2, Homo refers to homophobia, Trans refers to transphobia, LGBT refers to Non-anti-LGBT+ content, and Total refers to the total number of the data. From table 2 we can see that after balancing operation the datasets become balanced with equal numbers of each category.

4.3 Other Settings

We use the RoBERTa-base version in our experiments, and its output size is 768, so we set the input size of the first fully connected layer as 768 and the output size as 64, set the input size of the second fully connected layer as 64, and the output size as 3, the number of the categories of the dataset. We set the batch size of the data inputted to the model as 4 and trained our model with a 1e-5 learning rate.

Table 3: Details of results on test datasets

Language	Mac-F1 Score	Rank
English	0.57	1
Tamil	0.75	5
Tamil-English	0.53	6

5 Results

We use the RoBERTa -based approach to train the model on the training datasets of the task, validate the model on the development set, and use the trained model to predict the label of the test set. Repeated experiments with different epoch values, we found that when the epoch is 8, the trained model has the best validation results on the development set. So we train our model for 8 iterations, predict the labels of the test set with the model, and submitted the run results in all of the three languages. Details of results on test data sets and ranks are shown in table 3:

As we can see from table 3, our model achieved good results on the English subtask but the results of Tamil and English-Tamil subtasks are not so good. We use the same approach for the three subtasks but with different RoBERTa versions, RoBERTa-base for English subtask, and Tamil-RoBERTa for Tamil and Tamil-English subtasks. So we think that if just choose a suitable pre-trained language model, our approach will be effective for homophobia/transphobia detection task, and we also think that RoBERTa-base is suitable for English subtask, but Tamil-RoBERTa isn't very suitable for Tamil and Tamil-English subtasks.

From table3 we also can see that the f1-scores of Tamil and Tamil-English subtasks are 0.75 and 0.53 respectively, with a wide difference. To address the reason for the problem, we averaged the f1-scores of all teams for the two subtasks separately, then subtracted the average of f1-scores of the Tamil-English subtask from the average of f1-scores of the Tamil subtask, the difference between the two is 0.18. Then we subtracted the f1-score of our model on the Tamil-English subtask from the f1-score of our model on the Tamil subtask, the difference between the two is 0.22. So far we found that the f1-score of the Tamil subtask is generally higher than the f1-score of the Tamil-English subtask, so we think that the difference between the f1-scores of the two subtasks is related to the features of the datasets for Tamil and Tamil-English subtasks.

Table 4: Details of results on test datasets

Language	Mac-F1 Score
English	0.32
Tamil	0.29
Tamil-English	0.3145

To make a comparison between the test results of the models trained on balanced data sets (balanced by using RandomOverSampeler class) and the test results of the models trained on the original unbalanced train data sets after the competition. We train the models using the original unbalanced train data sets and test them

by using test data sets with labels. The results are shown in Table 4.

By comparing table 3 and table 4, we can find that there are big differences between the results. The test results of the models trained on balanced train data sets are much better than the test results of the models trained on the original unbalanced train data sets. Based on this, we can get the conclusion that balancing the train data sets by using RandomOverSampeler class is very effective and important in our experiments. Balancing the train data sets greatly improved the performance of the models.

6 Conclusion

We use a RoBERTa-based approach for homophobia/transphobia detection tasks and achieved better results in our experiments. Although the results on the Tamil and Tamil-English subtasks are not so ideal, the results on the English subtask show that our approach is effective for the homophobia/transphobia detection tasks. Although this competition has come to an end, there are still some directions we can continue to study in the future. For example, we can use prompt learning to process this task, by converting this task into cloze form and designing reasonable templates and verbalizers, we can fully make use of the knowledge of pre-trained language models, and may get better results. Besides, by converting the homophobia/ transphobia detection task into a text generation task, and then using text generation models like GPT (Radford et al.,2019), T5 (Raffel et al.,2020) to solve the task, We may get unexpected results. In future research, we will continue to study the directions mentioned above, and strive to achieve better homophobia/ transphobia detection performance.

References

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *Pytorch: An imperative style, high-performance deep learning library*. arXiv:1912.01703 [cs.LG].

Chakravarthi, B.R., Priyadharshini, R., Ponnusamy, R., Kumaresan, P.K., Sampath, K., Thenmozhi, D., Thangasamy, S., Nallathambi, R. and McCrae, J.P., 2021. *Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments.* arXiv preprint arXiv:2109.00227.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,Wei Li, and Peter J. Liu. 2020. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* In Journal of Machine Learning Research.

Diederik P. Kingma and Jimmy Ba. 2017. *Adam: A method for stochastic optimization.* arXiv:1412.6980 [cs.LG].

Divyansh Singh. 2021. *Detection of Emotions in Hindi-English Code Mixed Text Data.* arXiv:2105.09226 [cs.CL]

Divyanshu Kakwani and Anoop Kunchukuttan and Satish Golla and Gokul N.C. and Avik Bhattacharyya and Mitesh M. Khapra and Pratyush Kumar. 2020. *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages.* Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1–11.4948-4961.

Fernando Barrag an-Medero, David Perez-Jorge.2020. *Combating homophobia, lesbophobia, biphobia and transphobia:A liberating and subversive educational alternative for desires.* Heliyon 6 (2020) e05225.

Gamez-Guadix , Daniel Incera.2021. *Homophobia is online: Sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents.* Computers in Human Behavior 119 (2021) 106728.

Guillaume Lemaitre, Fernando Nogueira, Christos K. Aridas.2017. *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning.* Journal of Machine Learning Research, pages 1-5.

Ilya Loshchilov and Frank Hutter. 2019. *Decoupled weight decay regularization.* arXiv:1711.05101 [cs.LG].

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.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. 2019. *Language models are unsupervised multitask learners.* OpenAI blog, 1(8):9, 2019.

Ravindra Nayak, Raviraj Joshi. 2021. *Contextual Hate Speech Detection in Code Mixed Text using Transformer Based Approaches*. arXiv:2110.09338 [cs.CL].

Telmo Pires, Eva Schlinger, Dan Garrette. 2019. *How Multilingual is Multilingual BERT?* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996-5001.

Thomas Wolf, Lysandre Debut, Victor Sanh, JulienChaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R´emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu,Teven Le Scao, Sylvain Gugger, Mariama Drame,Quentin Lhoest, and Alexander M. Rush. 2020. *Huggingface's transformers: State-of-the-art natural language processing.* arXiv:1910.03771 [cs.CL].

Yezhe Lin, Hui Xie, Zimo Huang, Quan Zhang, Amanda Wilson, Jiaojiao Hou, Xudong Zhao, Yuanyuan Wang, Bailin Pan, Ye Liu, Meng Han, Runsen Chen. 2021. *The mental health of transgender and gender non-conforming people in China: a systematic review.* Lancet Public Health 2021;6: e9, pages 54–69.

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