IDIAP Submission@LT-EDI-ACL2022: Homophobia/Transphobia Detection in social media comments

Muskaan Singh, Petr Motlicek
IDIAP Research Institute,
Martigny, Swizerland
(msingh,petr.motlicek)@idiap.ch

Abstract

The increased expansion of abusive content on social media platforms negatively affects online users. Transphobic/homophobic content indicates hatred comments for lesbian, gay, transgender, or bisexual people. It leads to offensive speech and causes severe social problems that can make online platforms toxic and unpleasant to LGBT+people, endeavoring to eliminate equality, diversity, and inclusion. In this paper, we present our classification system; given comments, it predicts whether or not it contains any form of homophobia/transphobia with a Zero-Shot learning framework. Our system submission achieved 0.40, 0.85, 0.89 F1-score for Tamil and Tamil-English, English with (1^{st}, 1^{st},8^{th}) ranks respectively. We release our codebase here [1].

1 Introduction and Related Work

Homophobic/Transphobic (Diefendorf and Bridges, 2020; Giametta and Havkin, 2021) content on social media intends to harm Lesbian, Gay, Bi-sexual (LGB) (with labels such as *'fag', 'homo'* or denigrating phrases such as *'don't be a homo,' 'that's so gay'*) (Szymanski et al., 2008; Poteat and Rivers, 2010; Graham et al., 2011; Fraïssé and Barrientos, 2016).

It is a type of abuse that involves physical violence such as killing, maiming, beating, or explicit sexual violence such as rape, molestation, penetration, or an invasion of privacy by disclosing personal information.

Some of the example phrases include *"Gays deserve to be shot dead," "Someone should rape that lesbo to turn her into straight," "Gays should be stoned," "You lesbos, I know where you live, I will visit you tonight," "beat the fag out of him," "You should kill yourself"*.

[1] https://github.com/Muskaan-Singh/Homophobia-and-Transphobia-ACL-Submission.git

Social media has provided the freedom to express their views and thoughts on anything (Gkotsis et al., 2016; Wang et al., 2019), leading to unpleasant things on the internet (Zampieri et al., 2019).

Online offensive language has been identified as a worldwide phenomenon diffused throughout social media platforms such as Facebook, YouTube, and Twitter during the last decade (Gao et al., 2020).

It is even more distressing for Lesbian, Gay, Bisexual, Transgender, and other (LGBT+) vulnerable individuals (Díaz-Torres et al., 2020). Because of who they love, how they appear, or who they are, LGBT+ people all across the globe are subjected to violence and inequity, as well as torture and even execution (Barrientos et al., 2010; Schneider and Dimito, 2010). Sexual orientation and gender identity are essential components of our identities that should never be discriminated against or abused (Thurlow, 2001). However, in many countries, being identified as LGBT+ will cost lives, so the vulnerable individual goes to social media to get support or share their stories to find similar people (Adkins et al., 2018; Han et al., 2019). Identifying such information from social media would eliminate the severe societal problem and prevent formulating online platforms toxic unpleasant to LGBT+ people while also attempting to eliminate equality, diversity, and inclusion.

There are many rules and regulations to protect LGB persons, but they omit protection based on gender identity or expression or transgender adolescent experiences (McGuire et al., 2010; Hatchel et al., 2019).

Lack of annotated data has restrained the research on homophobic and transphobic detection. (Wu and Hsieh, 2017) find the linguistic behavior in LGBT+ for the Chinese language. The research experiments present the traditional system's failure for complex dimensions to detect the gender from the text. (Ljubešić et al., 2020) curated lexicons in

Table 1: Dataset Statics for training, development and test sets for English, Tamil and Tamil-English

	Train	Dev	Test
English	3164	792	990
Tamil	2662	666	833
Tamil-English	3861	966	1207

Croatian, Dutch, and Slovene for emotions. Further, the lexicons map the social text for migrants and LGBT+.

2 Shared Task Description

In the shared task, participants are provided with comments extracted from social media [2]. The challenge was to predict whether or not it contains any form of homophobia/transphobia detection. The participants are provided with a seed data (Chakravarthi et al., 2021), sampled as in Table: 1 respectively. The comments are manually annotated to show whether the text contains homophobia/transphobia. We also did reports data distribution across *Non-anti-LGBT+ content, Homophobic, Transphobic*, for all the languages in Table 2. Some examples for the *Non-anti-LGBT+ content, Homophobic, Transphobic* comments are presented in Table 3. In addition, it also provided a baseline code with machine learning algorithms (Multinomial Naive Bayes, SVM, KNN, Logistic Regression, and Decision trees).

3 Proposed Methodology

This section presents the proposed methodology for classifying Non-anti-LGBT+ content, Homophobic, and Transphobic content from social media posts. Initially, we preprocess the comments for special characters, stopwords, emojis, and punctuation removal using NLTK library (Loper and Bird, 2002). Further, we extract the features, tokenize all the sentences and map the tokens to their word IDs. For every sentence in the dataset, we follow a series of steps (i) tokenize the sentences (ii) prepend the [CLS] token to the start (iii) append the [SEP] token to the end (iv) map the token to their IDs (v) pad or truncate the sentenced to max length (vi) mapping of attention masks for [PAD] tokens. We padded and truncated the max_length=30. The generated sequence sentences are passed for encoding with its attention mask (simply differentiating padding from non-padding). Afterward, we feed these embeddings for pretraining the XLM ROBERTA (Conneau et al., 2019). It significantly aims at cross-lingual transfer tasks for pre-trained multilingual language models. The model performs exceptionally well on low resource languages at a scale. The empirical analysis presents positive transfer and capacity delusion. Further, the model also allows multilingual modeling without sacrificing per-language performance. It has shown competitive results with strong monolingual models on GLUE. After the pretraining, we fine-tune the model in the English language, and finally, we test on Tamil and Tamil-English languages.

[2] https://sites.google.com/view/lt-edi-2022

3.0.1 Experimental Setup

We use V1 100 GPU with 53GB RAM alongside 8 CPU cores for the experimental setup. We divide the entire dataset in 90:10 for train and validation of 8 batches, with learning rate (1e-5) and Adam optimizer (Kingma and Ba, 2014) with epsilon (1e-8). We feed a seed_val of 42. For calculating the training loss over all the batches, we use gradient descents (Ruder, 2016) with clipping the norm to 1.0 to avoid exploding gradient problem.

4 Results

We test our model for the dataset (Chakravarthi et al., 2021). The classification report for our proposed and the top-performing model over the test set can be seen in Table 7. The proposed model has proved itself remarkable by achieving 0.40, 0.85, 0.89 F-score with 1^{st}, 1^{st},8^{th} rank for Tamil and Tamil-English, English respectively on the leaderboardhttps://competitions.codalab.org/competitions/36394#learn_the_details-results. We also report, analysis of our results in Table: 6 corresponding to *Non-anti-LGBT+ content, Homophobic, Transphobic* labels.

- For the English language, 0.94, 0.54 are the reported precision, and recall, which is relatively 0.01, 0.03 more than the average and 0.01, and 0.04 less than the best performing model respectively. The reported F1-Score is 0.40 of our proposed model which is 0.03 less than the average, and 0.21 less than the best-performing.

- For the Tamil language, 0.94, 0.88, and 0.85 are the reported Precision, Recall, and F1-score, relative, 0.09, 0.18, and 0.18 more than

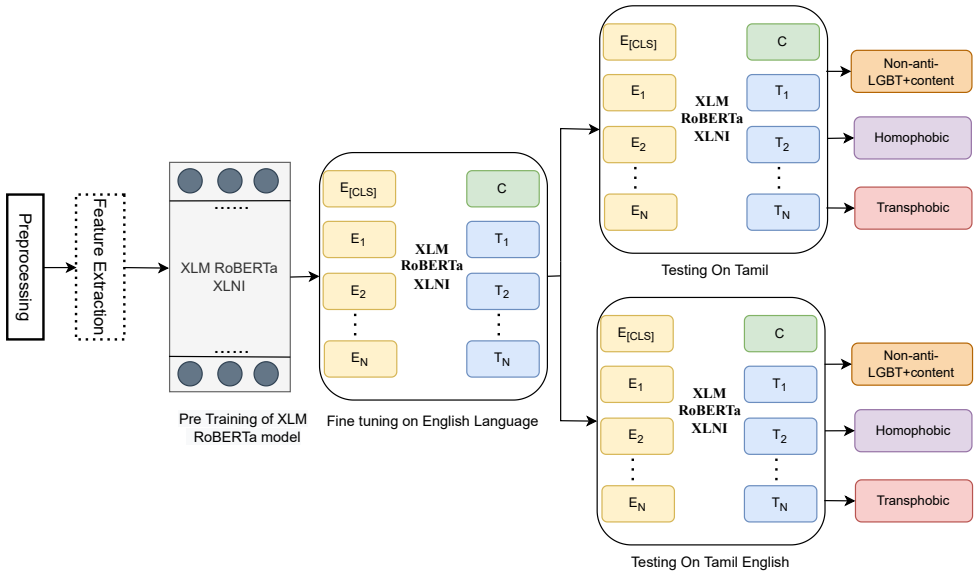

Figure 1: We predicted the labels using fine tuned XLM RoBERTa XLNI model.

Label	Language-wise distribution (Train + Dev)		
	English	**Tamil**	**English Tamil**
Non-anti-LGBT+ content	3733	2548	4300
Homophobic	215	588	377
Transphobic	8	192	150

Table 2: Data distribution for the Homophobia/Transphobia Detection in social media comments database.

Comment	**Label**
I support her, very smart ponnu	Non-anti-LGBT+ content
Stupid film there is no gays in the world these are all their imagine only	Homophobic
Hey seriously I thought She was a Transgender	Transphobic

Table 3: Examples for Non-anti-LGBT+ content, Homophobic, Transphobic in the Homophobia/Transphobia Detection in social media comments dataset.

Model	Precision		Recall		F1-Score	
	Macro	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Proposed model	0.94	0.94	0.88	0.94	0.85	0.94
Average score	0.85	0.84	0.70	0.85	0.67	0.85

Table 4: Comparison with the top-performing model results for Tamil.

Model	Precision		Recall		F1-Score	
	Macro	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Proposed model	0.63	0.89	0.60	0.89	0.61	0.89
Average score	0.54	0.87	0.52	0.87	0.51	0.86

Table 5: Comparison with the top-performing model results for Tamil-English.

Table 6: Prediction for Non-anti-LGBT+ content, Homophobic, Transphobic in the Homophobia/Transphobia Detection in social media comments dataset

Comment	Label
Best movie and people not understand relationship feeling I miss my life	Non-anti-LGBT+ content
gay culture does not suit the Indian culture. that's it.	Non-anti-LGBT+ content
Hormonal and psychological problem!!! Nothing more nothing less !!! Don't bring nature here and make it dirty !!!	Homophobic
This is even among animals and many other species. What country are you talking abt. Just foolish!	Homophobic

	Precision		Recall		F1-Score	
Model	**Macro**	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Top performing	0.95	0.94	0.58	0.95	0.61	0.94
Proposed model	0.94	0.92	0.54	0.94	0.40	0.92
Average score	0.93	0.92	0.51	0.93	0.43	0.92

Table 7: Comparison with the top-performing model results for English.

the average. As we were the best-performing submission, we did not report the scores and differences from our submission.

- For the Tamil English language, 0.63, 0.60, and 0.61 are the reported Precision, Recall, and F1-score, relative, 0.09, 0.08, and 0.10 more than the average. As we were the best-performing submission, we did not report the scores and differences from our submission.

The qualitative analysis predicted results are in Table 6. The true instances, "Best movie and people not understand relationship feeling I miss my life" and "This is even among animals and many other species. What country are you talking abt.Just foolish!" are labeled as Non-anti-LGBT+content and Homophobic, respectively. Unlike the other instances, these statements have precise negative/positive phrases that can help detect the sentiments. While the cases, "gay culture does not suit the Indian culture. that's it." labeled as Non-anti-LGBT+content, but it is a homophobic comment on reading the sentence. It indicates that the model focused more on words such as "gay" and "suit" rather than the entire meaning of the statement. "Hormonal and psychological problem!!! Nothing more nothing less !!!
Don't bring nature here and make it dirty !!! " instance is labeled as homophobic, but in our opinion, it is supporting the cause. It signifies that the model is more focused on the negative sentiments such as "fool" and "animals" rather than understanding the entire context of the comment.

5 Conclusion

In this paper, we present our classification system; given comments, it predicts whether or not it contains any form of homophobia/transphobia with a zero-shot learning framework. Our system submission achieved 0.40, 0.85, 0.89 F1-score for Tamil and Tamil-English, English with (1^{st}, 1^{st},8^{th}) ranks respectively. We also performed a qualitative analysis. The system performs precisely on negative/positive phrases such as "fool" and "animals" rather than understanding the entire context of the comment. We intend to work on a multi-task learning framework to handle different kinds of homophobic/transphobic by capturing context in the future. We also aim to detect multilingual homophobic/transphobic comments in the code-mixing scenarios.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

Victoria Adkins, Ellie Masters, Daniel Shumer, and Ellen Selkie. 2018. Exploring transgender adolescents' use of social media for support and health information seeking. *Journal of Adolescent Health*, 62(2):S44.

Jaime Barrientos, Jimena Silva, Susan Catalán, Fabiola Gómez, and Jimena Longueira. 2010. Discrimination and victimization: parade for lesbian, gay, bisexual, and transgender (lgbt) pride, in chile. *Journal of homosexuality*, 57(6):760–775.

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

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*.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes, Juan Aguilera, and Luis Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a mexican spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136.

Sarah Diefendorf and Tristan Bridges. 2020. On the enduring relationship between masculinity and homophobia. *Sexualities*, 23(7):1264–1284.

Christèle Fraïssé and Jaime Barrientos. 2016. The concept of homophobia: A psychosocial perspective. *Sexologies*, 25(4):e65–e69.

Junling Gao, Pinpin Zheng, Yingnan Jia, Hao Chen, Yimeng Mao, Suhong Chen, Yi Wang, Hua Fu, and Junming Dai. 2020. Mental health problems and social media exposure during covid-19 outbreak. *Plos one*, 15(4):e0231924.

Calogero Giametta and Shira Havkin. 2021. Mapping homo/transphobia. *ACME: An International Journal for Critical Geographies*, 20(1):99–119.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 63–73.

Robert Graham, Bobbie Berkowitz, Robert Blum, Walter Bockting, Judith Bradford, Brian de Vries, and Harvey Makadon. 2011. The health of lesbian, gay, bisexual, and transgender people: Building a foundation for better understanding. *Washington, DC: Institute of Medicine*, 10:13128.

Xi Han, Wenting Han, Jiabin Qu, Bei Li, and Qinghua Zhu. 2019. What happens online stays online?——social media dependency, online support behavior and offline effects for lgbt. *Computers in Human Behavior*, 93:91–98.

Tyler Hatchel, Gabriel J Merrin, Espelage, and Dorothy. 2019. Peer victimization and suicidality among lgbtq youth: The roles of school belonging, self-compassion, and parental support. *Journal of LGBT Youth*, 16(2):134–156.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The lilah emotion lexicon of croatian, dutch and slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 153–157.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Jenifer K McGuire, Charles R Anderson, Russell B Toomey, and Stephen T Russell. 2010. School climate for transgender youth: A mixed method investigation of student experiences and school responses. *Journal of youth and adolescence*, 39(10):1175–1188.

V Paul Poteat and Ian Rivers. 2010. The use of homophobic language across bullying roles during adolescence. *Journal of Applied Developmental Psychology*, 31(2):166–172.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Margaret S Schneider and Anne Dimito. 2010. Factors influencing the career and academic choices of lesbian, gay, bisexual, and transgender people. *Journal of homosexuality*, 57(10):1355–1369.

Dawn M Szymanski, Susan Kashubeck-West, and Jill Meyer. 2008. Internalized heterosexism: A historical and theoretical overview. *The Counseling Psychologist*, 36(4):510–524.

Crispin Thurlow. 2001. Naming the "outsider within": Homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. *Journal of adolescence*, 24(1):25–38.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552.

Hsiao-Han Wu and Shu-Kai Hsieh. 2017. Exploring lavender tongue from social media texts [in chinese]. In *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, pages 68–80.

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). *arXiv preprint arXiv:1903.08983*.