JudithJeyafreeda@LT-EDI: Using GPT model for recognition of Homophobia/Transphobia detection from social media

Judith Jeyafreeda Andrew
University of Manchester
Oxford Rd, Manchester M13 9PL
Manchester,UK
judithjeyafreeda@gmail.com

Abstract

Homophobia and Transphobia is defined as hatred or discomfort towards Gay, Lesbian, Transgender or Bisexual people. With the increase in social media, communication has become free and easy. This also means that people can also express hatred and discomfort towards others. Studies have shown that these can cause mental health issues. Thus detection and masking/removal of these comments from the social media platforms can help with understanding and improving the mental health of LGBTQ+ people. In this paper, GPT2 is used to detect homophobic and/or transphobic comments in social media comments. The comments used in this paper are from five (English, Spanish, Tamil, Malayalam and Hindi) languages. The results show that detecting comments in English language is easier when compared to the other languages.

1 Introduction

Homophobic and/or Transphobic comments is a form of Hate Speech directed towards LGBTQ+ community. With the increase in internet and use of social media, the use of derogatory comments have increased considerably. These comments cause mental health issues for a lot of people within the LGBTQ+ community (Chakravarthi et al., 2022a,b; Chakravarthi, 2023). Thus identification of these comments is necessary to improve the well being of the community. This is a specific case of offensive language or Hate speech detection.

The task in this paper, is to identify and classify text into 3 classes (sub task 1) or 7 classes (sub task 2) [detailed in section 2]. The language concerned in this task are English, Tamil, Malayalam, Hindi and Spanish. Some text are code-mixed text. Code-mixing is the process of mixing more than one language in a text. Chakravarthi et al. (2020) and Chakravarthi et al. (2022c) have developed a dataset and methods for sentiment analysis for code-mixed data for the Dravidian languages of Tamil and English. The task in this paper is a multi class classification problem. In this task, there are more than 2 predefined classes and each text can be placed in only one of the predefined class. Several multi class classification approaches have been proposed previously like in (Thavareesan and Mahesan, 2019), (Thavareesan and Mahesan, 2020a). However, considering the languages and context, all the methods might not be suitable for the task at hand. (Thavareesan and Mahesan, 2020b) have proposed a embedding for the language Tamil. Other forms of pre processing for Dravidian languages have been proposed by Ghanghor et al. (2021); Puranik et al. (2021); U Hegde et al. (2021); Yasaswini et al. (2021)

2 Task Description

In this task in Chakravarthi et al. (2022a), comments from social media from different languages are used for the classification. The task has two sub tasks. In the first subtask, the comments are from five languages (English, Spanish, Tamil, Malayalam and Hindi) with 3 labels (Non-anti-LGBT+ content, Homophobia and Transphobia). The second subtask has comments from 3 languages (English,Tamil and Malayalam) with 7 labels (Counter-speech, Homophobic-derogation, Homophobic-Threatening, Hope-Speech, Transphobic-derogation, Transphobic-Threatening, None-of-the-above). Tables 1 and 2 shows the number of comments in each set for the different languages and different sub tasks.

3 Related Work

Detection and Classification homophobic and transphobic comments can be considered as a specific case of Hate Speech detection. There has been

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 78–82, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_011

Language	Train	Dev	Test
English	3164	792	990
Tamil	2662	666	831
Malayalam	3114	1211	864
Hindi	2560	318	321
Spanish	850	236	500

Table 1: Data statistics for Sub Task 1

Language	Train	Dev	Test
English	3149	792	990
Tamil	2662	666	833
Malayalam	3114	1213	866

Table 2: Data statistics for Sub Task 2

several works done in the field of hate speech or offensive language detection detection. Within this field several work has been done. Machine Learning methods have been commonly used for classification in several works such as (Yin et al., 2009), (Dadvar et al., 2013), (ming Xu et al.), (Razavi et al., 2010), (Spertus, 1997). These work focus on cyber bullying, where a machine learning model is used to classify text into specified categories such that cyber bullying can be detected and reported. (Rodríguez-Ibánez et al., 2023) proposes a comprehensive review for the sentiment analysis methods applied on social media data. The authors review both academic and industrial tools that have been developed for the purpose of sentiment analysis of social media texts.

Recent efforts on classification of offensive text involve the use of Neural Networks. Within this context, (Risch et al., 2020) compare four models: an interpretable machine learning model (naive Bayes), a model-agnostic explanation method (LIME), a model-based explanation method (LRP), and a self-explanatory model (LSTM with an attention mechanism), showing that complex models perform better than simpler ones.

Most work done within this area focuses on the English Language, however there are language processing challenges when different languages are used. (Chakravarthi, 2022b; Kumaresan et al., 2022; Chakravarthi, 2022a) presents an improvement of word sense translation for under-resourced languages. (Jeyafreeda, 2020) proposed a Multiclass Classification method, where several Machine Learning algorithms have been adapted to the task of sentiment analysis and based on the accuracy of the algorithms on the development set the best suited technique is chosen for the language and the task. (Andrew, 2021) suggests few machine language approaches to classify texts from Code-mixed Dravidian Languages. (Andrew, 2022) uses a CNN approach for the classification of emotion in YouTube comments for the dravidian language of Tamil. In this paper, the data from various languages are pre-processed with using methods described in (Andrew, 2021) and (Andrew, 2022). This is then used along with a GPT model for classification.

4 Proposed System

In this work GPT2 is used for classification of Homophobic and Transphobic comments. The model is finetuned on the training dataset for each task and every language. For languages other than English, the text is replaced with the IPA equivalent, this approach has been inspired from (Andrew, 2021) and (Andrew, 2022). The categories are in English language, thus IPA equivalent character need not be substituted.

Pre-processing: Similar to (Andrew, 2022), a few steps of pre-processing is performed to get the accurate representation of the text.

This involves the following:

- Texts from languages other than English into IPA text equivalents. The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin script. This is performed using the anyascii package in Python.
- The emojis are substituted with the words of the emotion they represent like happy, sad, excited etc.
- The tokenizer from the pretrained GPT2 model is used for tokenization of the transformed text.

GPT2 GPT, Generative Pre trained models, is a neural network based architecture which uses transformers. These use a self-attention mechanism allowing to focus on different parts of the input text during the various stages on processing. GPT-2 model has 1.5 billion parameters and has been trained on 8 million web pages in a self-supervised fashion. (Radford et al., 2019) provides a detailed description of the model. The inputs are sequences of continuous text of a certain length

and the targets are the same sequence, shifted one token (word or piece of word) to the right. The model uses internally a mask-mechanism to make sure the predictions for the token i only uses the inputs from 1 to i but not the future tokens. This allows the model to learn the inner representation of the language, which can then be used to extract features for downstream tasks.

GPT2 for classification: Python has a range of packages that allow the use of GPT models such as Hugging Face's Transformers, NLTK, and TextBlob. In this paper, this python package is used for classification of text into classes of sentiments. The training data for each language is used to fine tune these models with the different classes (varying for the two sub classes) and for each language.

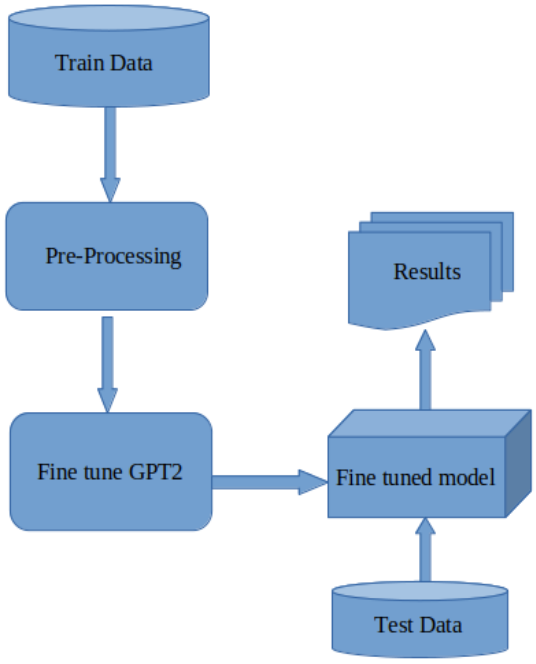

Figure 1: Process flow

5 Evaluation

The performance of the classification system is measured in terms of macro averaged Precision, macro averaged Recall and macro averaged F-Score across all the classes (for both sub tasks). The Scikit-learn [1] package is used for this purpose.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

6 Results

Language	Weighted F1
English	0.90
Tamil	0.27
Malayalam	0.25
Hindi	0.02
Spanish	0.0

Table 3: Results of Sub Task 1

Language	Weighted F1
English	0.23
Tamil	0.65
Malayalam	0.06

Table 4: Results of Sub Task 2

Tables 3 and 4 show the results of the sub tasks 1 and 2 respectively. To recap, the sub task 1 is to classify the text into 3 classes (Non-anti-LGBT+ content, Homophobia and Transphobia) and the sub task 2 is to classify the text into 7 classes (Counter-speech, Homophobic-derogation, Homophobic-Threatening, Hope-Speech, Transphobic-derogation, Transphobic-Threatening, None-of-the-above). Although the models are specifically fine tuned for each languages and sub tasks, some fine tuned models perform better than the others. From Table 3, it can be noted that the best results of the model are for the English language, this is obvious considering the model has been trained initially with English texts. However, table 4 shows that the model achieves better performance the Tamil language with the F1 score of 0.65, while for the English language the score is 0.23. This is an interesting result, considering that the Tamil Language texts have been replaced with IPA format text while the English language text went through no such pre processing. This could be because of the increase in the number of classes for classification. The most common class in the training data for the tamil language was "None-of-the-above", while the English language had several texts in each classes. The model was not a success for the Spanish and Hindi language, as seen from 3. For the other languages, the models achieve an average of 0.20 for F1-score. This is not the best results. This indicates that several improvements need to

be made to adapt models into other languages.

The replacement of text with IPA characters have been efficient for some machine learning models (Andrew, 2021) and (Andrew, 2022), however it might not be the best representation for a transformer based model. Choosing to use a different embedding system might prove to be more efficient, such as (Thavareesan and Mahesan, 2020b) for the Tamil language. A tokenizer designed for specific languages can be used in place of the GPT2 pretrained tokenizer. Improving the balance of classes in the training set could help in better classification of the test set.

References

Judith Jeyafreeda Andrew. 2021. JudithJeyafreedaAndrew@DravidianLangTech-EACL2021:offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv. Association for Computational Linguistics.

Judith Jeyafreeda Andrew. 2022. JudithJeyafreedaAndrew@TamilNLP-ACL2022:CNN for emotion analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 58–63, Dublin, Ireland. Association for Computational Linguistics.

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

Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in English and Dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in Youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

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, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022c. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. pages pp 693–696.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics.

Judith Jeyafreeda. 2020. JudithJeyafreeda@DravidianCodeMix-FIRE2020:Sentiment Analysis of YouTube Comments for Dravidian Languages.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@LT-EDI-EACL2021-hope speech detection: There is always hope in transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 98–106, Kyiv. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using

multi-level classification. In *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. 2023. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based part of speech tagging in tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. UVCE-IIITT@DravidianLangTech-EACL2021: Tamil troll meme classification: You need to pay more attention. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–186, Kyiv. Association for Computational Linguistics.

Jun ming Xu, Kwang sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian Davison, April Edwards, and Lynne Edwards. 2009. Detection of harassment on web 2.0.