MUCS@LT-EDI-2024: Learning Approaches to Empower Homophobic/Transphobic Comment Identification

Sonali[a], **Nethravathi Gidnakanala**[b], **Raksha G**[c], **Kavya G**[d], **Asha Hegde**[e], **H L Shashirekha**[f]
Department of Computer Science, Mangalore University, Mangalore, Karnataka, India
{[a]sonalikulal417, [b]nethravathig749, [c]rakshagmangalore}@gmail.com,
{[d]kavyamujk, [e]hegdekasha}@gmail.com, [f]hlsrekha@mangaloreuniversity.ac.in

Abstract

Homophobic/Transphobic (H/T) content includes hatred and discriminatory comments directed at Lesbian, Gay, Bisexual, Transgender, Queer (LGBTQ) individuals on social media platforms. As this unfavourable perception towards LGBTQ individuals may affect them physically and mentally, it is necessary to detect H/T content on social media. This demands automated tools to identify and address H/T content. In view of this, in this paper, we - team MUCS describe the learning models submitted to "Homophobia/Transphobia Detection in social media comments:LT-EDI@EACL 2024" shared task at European Chapter of the Association for Computational Linguistics (EACL) 2024. The learning models: i) Homo_Ensemble - an ensemble of Machine Learning (ML) algorithms trained with Term Frequency-Inverse Document Frequency (TF-IDF) of syllable n-grams in the range (1, 3), ii) Homo_TL - a model based on Transfer Learning (TL) approach with Bidirectional Encoder Representations from Transformers (BERT) models, iii) Homo_probfuse - an ensemble of ML classifiers with soft voting trained using sentence embeddings (except for Hindi), and iv) Homo_FSL - Few-Shot Learning (FSL) models using Sentence Transformer (ST) (only for Tulu), are proposed to detect H/T content in the given languages. Among the models submitted to the shared task, the models that performed better for each language include: i) Homo_Ensemble model obtained macro F1 score of 0.95 securing 4th rank for Telugu language, ii) Homo_TL model obtained macro F1 scores of 0.49, 0.53, 0.45, 0.94, and 0.95 securing 2nd, 2nd, 1st, 1st, and 4th ranks for English, Marathi, Hindi, Kannada, and Gujarathi languages, respectively, iii) Homo_probfuse model obtained macro F1 scores of 0.86, 0.87, and 0.53 securing 2nd, 6th, and 2nd ranks for Tamil, Malayalam, and Spanish languages respectively, and iv) Homo_FSL model obtained a macro F1 score of 0.62 securing 2nd rank for Tulu dataset.

1 Introduction

Homophobia and Transphobia are the two terms that refer to negative attitude towards the homosexual and transsexual people like LGBTQ. These attitudes are expressed in terms of H/T comments, insults, and discriminatory language on social media platforms (Chakravarthi, 2023; Hegde et al., 2023a). This unfavourable perceptions towards homosexual and transsexual people can have a very negative effect which can exacerbate mental health issues and give them a sense of helplessness and fear (Chakravarthi et al., 2022). Hence, there is a need to develop automated tools to detect H/T content to maintain healthy social media platforms.

In a multilingual country like India, people prefer to blend English words or sub-words with their native language creating code-mixed texts (Chakravarthi, 2023). The intricate nature of code-mixed text introduces additional complexities, where words or sub-words from different languages may be combined in different ways lacking grammatical rules, making it challenging to establish consistent patterns for classification. The H/T content available on social media may also be in code-mixed form (Hegde and Shashirekha, 2022b).

To address the challenges of H/T content detection in social media text, in this paper, we - team MUCS, describe the models submitted to "Homophobia/Transphobia Detection in social media comments:LT-EDI@EACL 2024" shared task[1] (Chakravarthi et al., 2024; Kumaresan et al., 2023). While the shared task is modeled as a multi-class text classification problem for H/T content detection in English, Hindi, Tamil, Telugu, Kannada, Gujarathi, Malayalam, Marathi, and Spanish languages, by employing ML and TL approaches, H/T content detection in Tulu is modeled as binary text classification problem using FSL approach.

The rest of the paper is structured as follows:

[1]https://codalab.lisn.upsaclay.fr/competitions/16056

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 288–293

Section 2 contains related works and Section 3 explains the methodology. Section 4 describes the experiments and results and the paper concludes in Section 5 with future work.

2 Related Work

Researchers have explored different approaches to detect H/T content on social media platforms. Description of few research works that are carried out to perform similar tasks are given below:
Ashraf et al. (2022) presented ML models (Support Vector Machine (SVM), Random Forest (RF), Passive Aggressive Classifier, Gaussian Naïve Bayes (GNB), Multi-Layer Perceptron) trained with TF-IDF of word bigrams, for H/T content detection in English, Tamil and Tamil-English. Out of their proposed models, SVM classifier outperformed the other classifiers with weighted F1 scores of 0.91, 0.92, and 0.88 for English, Tamil, and Tamil-English respectively. Singh and Motlicek (2022) fine-tuned Cross Lingual Language Models Robustly Optimised BERT (XLM-RoBERTa) model in the Zero-Shot learning framework for detecting H/T contents in English, Tamil, and Tamil-English texts. Their proposed methodology obtained weighted F1 scores of 0.92, 0.94, and 0.89 for English, Tamil, and Tamil-English languages respectively.

Pranith et al. (2022) presented TL based approach with two different BERT variants (IndicBERT and LaBSE (Language-Agnostic BERT Sentence Embedding)) for H/T content detection in English, Malayalam, Tamil-English, and Tamil languages. Their proposed LaBSE model obtained weighted F1 score of 0.46 for English language and IndicBERT model obtained weighted F1 scores of 0.54, 0.39, and 0.28 for Malayalam, Tamil-English, and Tamil languages respectively. Chanda et al. (2022) fine-tuned Multilingual BERT (mBERT) model for detecting sentiment and homophobia content in Malayalam and Kannada code-mixed texts and obtained macro F1 scores of 0.72 and 0.66 for Malayalam and Kannada code-mixed texts respectively. Nozza et al. (2022) fine-tuned different Large Language Models (LLMs) (BERT, RoBERTa, HateBERT) and ensemble modeling with majority voting to combine different fine-tuned LLMs. To handle the class imbalance they performed data augmentation by collecting external dataset to include more H/T instances for the detection of H/T content in English dataset. Their proposed ensemble model outperformed other models with a weighted F1 score of 0.94 for English dataset.

Though there are several models to identify H/T content in social media text, there is still scope for developing models for H/T content detection in low-resource languages like Tamil, Malayalam, Telugu, Tulu, etc., as these languages are not much explored in the realm of code-mixed content.

3 Methodology

The proposed methodology includes implementation of a wide range of learning models including ML, TL, and FSL approaches for identifying H/T content in the datasets provided by the organizers of the shared task. Pre-processing techniques are applied commonly to all the datasets. As the datasets are imbalanced, resampling techniques are used to balance the datasets in some of the learning models. Pre-processing and Resampling steps are explained below:

- **Pre-processing** - play an important role in text processing. Emojis are converted to the corresponding English text allowing them to be used as other text data followed by removing URLs, digits, and punctuation, as they do not contribute to text classification. Further, stopwords are removed using the corresponding references (English[2], Hindi[3], Tamil[4], Telugu[5], Kannada[6], Gujarathi[7], Marathi[8], and Spanish[9]).

- **Resampling** - When the number of instances in a labeled dataset for classification varies noticeably, the situation is referred to as data imbalance (Hegde et al., 2023b). This results in learning models becoming biased towards majority class, exhibiting poor performance for minority class. Resampling techniques are capable to resolve this biased training to

[2]https://www.nltk.org/search.html?q=stopwords
[3]https://github.com/stopwords-iso/stopwords-hi
[4]https://gist.github.com/arulrajnet e82a5a331f78a5cc9b6d372df13a919c
[5]https://github.com/Xangis/extra-stopwords/blob/master/telugu
[6]https://gist.github.com/MSDarshan91/f97c73435a3ab32a6638436231bf5616
[7]https://github.com/stopwords-iso/stopwords-gu/blob/master/stopwords-gu.txt
[8]https://github.com/stopwords-iso/stopwords-mr/blob/master/stopwords-mr.txt
[9]https://github.com/Alir3z4/stop-words/blob/master/spanish.txt

Language	Train set			Development set		
	None	Homophobia	Transphobia	None	Homophobia	Transphobia
English	2,978	179	7	748	42	2
Hindi	2,423	45	92	305	2	13
Tamil	2,064	453	145	507	118	41
Telugu	3,496	2,907	2,647	747	588	605
Kannada	4,463	2,765	2,835	955	585	617
Gujarathi	3,848	2,267	2,004	788	498	454
Malayalam	2,468	476	170	937	197	79
Marathi	2,572	551	377	541	129	80
Spanish	700	250	250	200	93	93
Tulu	542	188	-	-	-	-

Table 1: Statistics of the datasets

Language	Dev set	Test set
English	0.39	0.37
Hindi	0.33	0.33
Tamil	0.90	0.82
Telugu	0.96	**0.95**
Kannada	0.93	0.93
Gujarathi	0.95	0.99
Malayalam	0.93	0.94
Marathi	0.49	0.51
Spanish	0.78	0.51

Table 2: Performances of proposed Homo_Ensemble model in terms of macro F1 score

Language	Dev set	Test set
English	0.41	0.42
Tamil	0.85	**0.86**
Telugu	0.93	0.93
Kannada	0.32	0.54
Gujarathi	0.95	0.95
Malayalam	0.85	**0.87**
Marathi	0.54	0.52
Spanish	0.74	0.53

Table 3: Performances of the proposed Homo_probfuse model based on macro F1 score

some extent. Oversampling is a resampling technique that duplicates the samples belonging to the minority class and adds them to the Train set until it gets balanced. In this study, oversampling technique is used to balance English, Hindi, Tamil, Malayalam, Marathi, and Spanish datasets.

The description of the proposed learning models to identify H/T content in English, Hindi, Tamil, Telugu, Kannada, Gujarathi, Malayalam, Marathi, Spanish, and Tulu languages is given below:

3.1 Homo_Ensemble model

The proposed Homo_Ensemble model comprises of two modules: Feature Extraction and Classifier Construction as explained below:

3.1.1 Feature Extraction

Feature extraction is the process of extracting distinguishable features that can be used to train the learning models. Syllable representation provides meaningful tokens for Indian languages in native scripts. The given datasets (except English dataset) are syllabalized using IndicNLP[10] library. n-grams are a sequence of 'n' consecutive units, where the units can be characters, syllables or words. Syllable sequences in the range (1, 3) obtained from the given text are vectorized using TFIDFVectorizer[11] and the resultant feature vectors are used to train the classifiers.

3.1.2 Classifier Construction

Ensembling classifiers offers a potent method for overcoming individual classifier shortcomings by utilizing the strengths of other classifiers with the aim of improving the performance of a group of classifiers. This work ensembles ML classifiers (DT, SVM, NB, and Linear Support Vector Classifier (LSVC)) with hard voting.

3.2 Learning Models using Transformers

The proposed strategy of using transformers for classification is described below:

[10]https://indicnlp.ai4bharat.org/pages/home/

[11]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

3.2.1 Homo_TL model

TL approach involves utilizing knowledge acquired from one task to improve the performance of other but similar task. Instead of training a model from scratch for a new task, TL model leverages the knowledge using pre-trained models (Hegde and Shashirekha, 2022a). In this work, different BERT variants: Multilingual Bidirectional Encoder Representations from Transformers (mBERT)[12] (Devlin et al., 2018) (Hindi, Tamil, Malayalam, Marathi), gujarathi_sbert[13] (Deode et al., 2023; Joshi et al., 2022) (Gujarathi), KannadaSBERT-STS (kannada_sbert)[14] (Deode et al., 2023; Joshi et al., 2022) (Kannada), BERT[15] (Devlin et al., 2018) (English), Spanish BERT[16] (Cañete et al., 2020) (Spanish), and Homophobia_mBERT[17] (Telugu), are fine-tuned on the given Train sets. As the given Train sets are imbalanced, oversampling technique is used to balance the Train sets of English, Hindi, Tamil, Malayalam, Marathi, and Spanish language, before they are used to fine-tune for the intended task.

3.2.2 Homo_probfuse model

Soft voting is a type of ensemble method that involves assigning a probability score to each class for each model during ensembling. The final prediction is then determined by considering the probabilities of all the models. In this work, SVM and RF classifiers are trained using two Sentence Transformers (ST): mXLMR[18] (Reimers and Gurevych, 2019) and IndicSBERT-STS (indic_sbert)[19] (Deode et al., 2023)respectively, for the all datasets except Spanish, Hindi and Tulu languages. For Spanish language, Spanish BERT[20] and mXLMR are used to train SVM and RF classifiers respectively. The predictions of these classifiers are combined based on the maximum probability values. Additionally, the provided Train sets are oversampled before being trained on SVM and RF classifiers.

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

[13]l3cube-pune/gujarati-sentence-similarity-sbert

[14]l3cube-pune/kannada-sentence-similarity-sbert

[15]https://huggingface.co/bert-base-uncased

[16]https://huggingface.co/mrm8488/distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es

[17]https://huggingface.co/bitsanlp/Homophobia-Transphobia-v2-mBERT-EDA

[18]https://huggingface.co/sentence-transformers/stsb-xlm- r-multilingual

[19]l3cube-pune/indic-sentence-similarity-sbert

[20]https://huggingface.co/mrm8488/distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es

Model	Precision	Recall	Macro F1 score
ST_indic	0.61	0.68	0.61
ST_kan	0.62	0.70	**0.62**

Table 4: Performances of the proposed Homo_FSL models for Tulu Language

3.3 Homo_FSL model

ST framework of Python creates contextualised sentence embeddings for the given sentences. Few-shot and zero-shot approaches have received a great deal of interest in the research community due to the availability of ST and their untapped capacity to use them in resource-constrained domains (Girish et al., 2023). In view of this, Homo_FSL models are implemented to detect H/T content in Tulu text using two distinct ST models: indic_sbert and kannada_sbert for extracting sentence embeddings. The sentences in the given text are represented as sentence embeddings using the ST model and the mean embeddings of the sentence embeddings of a given text are obtained to train the ensemble of ML classifiers (LR, BernoulliNB (BNB), SVC, and RF) with hard voting.

4 Experiments and Results

Statistics of the dataset provided by the organizers (Chakravarthi et al., 2022, 2024, 2023) of the shared task for the detection of H/T contents in social media text for English, Hindi, Tamil, Telugu, Kannada, Gujarathi, Malayalam, Marathi, and Spanish are shown in Table 1. From Table 1, it is clear that the datasets provided are highly imbalanced. To overcome this, Homo_TL and Homo_probfuse models are experimented using resampled data by using oversampling method provided by the sklearn library[21]. Performances of the proposed Homo_Ensemble, Homo_probfuse, and Homo_FSL models are shown in the Tables 2, 3, and 4 respectively. Performances of Homo_TL model before and after oversampling are shown in the Table 5.

5 Conclusion and Future Work

This paper describes the models submitted by our team - MUCS, to the shared task "Homophobia/Transphobia Detection in social media comments: LT-EDI@EACL 2024" shared task at EACL 2024. Four distinct models: i) Homo_Ensemble

[21]https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html

Language	Development set		Test set	
	Before Oversampling	After Oversampling	Before Oversampling	After Oversampling
English	0.32	0.45	0.32	**0.49**
Hindi	0.33	0.41	0.33	**0.45**
Tamil	0.29	0.79	0.29	0.83
Telugu	0.95	-	0.95	-
Kannada	0.96	-	**0.94**	-
Gujarathi	0.96	-	**0.95**	-
Malayalam	0.49	0.89	0.49	0.91
Marathi	0.20	0.42	0.19	**0.53**
Spanish	0.82	0.43	0.49	0.42

Table 5: Performances of the proposed Homo_TL models before and after Oversampling

ii) Homo_TL and iii) Homo_probfuse are implemented to identify H/T content in all the given languages except Hindi and Tulu, and iv) Homo_FSL model is implemented only for Tulu dataset.

Among the models submitted to the shared task, only the models that performed better for each language are reported. Homo_Ensemble model obtained macro F1 score of 0.95 securing 4th rank for Telugu, Homo_TL model obtained macro F1 scores of 0.49, 0.53, 0.45, 0.94, and 0.95 securing 2nd, 2nd, 1st, 1st, and 4th ranks for English, Marathi, Hindi, Kannada and Gujarathi languages respectively and proposed Homo_probfuse model obtained macro F1 scores of 0.86, 0.87, and 0.53 securing 2nd, 6th, and 2nd ranks for Tamil, Malayalam, and Spanish languages respectively. Homo_FSL model trained using kannada_sbert obtained macro F1 score of 0.62 securing 2nd rank in the shared task for Tulu language. In the future, data augmentation methods for managing unbalanced classes using efficient feature extraction methods will be explored.

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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Bharathi Raja Chakravarthi. 2023. Detection of Homophobia and Transphobia in YouTube Comments. In *International Journal of Data Science and Analytics*, pages 1–20. Springer.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. How can we Detect Homophobia and Transphobia? Experiments in a Multilingual Code-mixed Setting for Social Media Governance. In *International Journal of Information Management Data Insights*, volume 2, page 100119. Elsevier.

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, Malliga Subramanian, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, and Nitesh Jindal. 2023. Overview of Second Shared Task on Homophobia and Transphobia Detection in English, Spanish, Hindi, Tamil, and Malayalam. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Supriya Chanda, Anshika Mishra, and Sukomal Pal. 2022. Sentiment Analysis and Homophobia Detection of Code-Mixed Dravidian Languages Leveraging Pre-trained Model and Word-level Language Tag. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A Simple Approach for Learning Cross-Lingual Sentence Representations using Multilingual BERT. In *arXiv preprint arXiv:2304.11434*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *CoRR*, volume abs/1810.04805.

Kavya Girish, A Hegdev, Fazlourrahman Balouchzahi, and SH Lakshmaiah. 2023. Profiling Cryptocurrency Influencers with Sentence Transformers. In *Working Notes of CLEF*.

Asha Hegde, G Kavya, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023a. MUCS@ LT-EDI2023: Homophobic/Transphobic Content Detection in Social Media Text using mBERT. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 287–294.

Asha Hegde, G Kavya, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023b. MUCS@ LT-EDI2023: Homophobic/Transphobic Content Detection in Social Media Text using mBERT. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 287–294.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022a. Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022b. Leveraging Dynamic Meta Embedding for Sentiment Analysis and Detection of Homophobic. In *Transphobic Content in Code-mixed Dravidian Languages*.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2022. L3Cube-MahaSBERT and HindSBERT: Sentence BERT Models and Benchmarking BERT Sentence Representations for Hindi and Marathi. In *arXiv preprint arXiv:2211.11187*.

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. In *Natural Language Processing Journal*, page 100041. Elsevier.

Debora Nozza et al. 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*. Association for Computational Linguistics.

P Pranith, V Samhita, D Sarath, and Durairaj Thenmozhi. 2022. Homophobia and Transphobia Detection of YouTube Comments in Code-Mixed Dravidian Languages using Deep learning.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Muskaan Singh and Petr Motlicek. 2022. IDIAP Submission@ LT-EDI-ACL2022: Homophobia/Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 356–361.