Tercet@LT-EDI: Homophobia/Transphobia Detection in social media comment

S Shwetha
SSN College of Engineering
shwetha2210210@ssn.edu.in

Samyuktaa Sivakumar
SSN College of Engineering
samyuktaa2210189@ssn.edu.in

Priyadharshini T
SSN College of Engineering
priyadharshini2210228@ssn.edu.in

Durairaj Thenmozhi
SSN College of Engineering
thenid@ssn.edu.in

B. Bharathi
SSN College of Engineering
bharathib@ssn.edu.in

Krithika S
SSN College of Engineering
krithika2010039@ssn.edu.in

Abstract

The advent of social media platforms has revolutionized the way we interact, share, learn, express, and build our views and ideas. One major challenge of social media is hate speech. Homophobia and transphobia encompass a range of negative attitudes and feelings towards people based on their sexual orientation or gender identity. Homophobia refers to the fear, hatred, or prejudice against homosexuality, while transphobia involves discrimination against transgender individuals. Natural Language Processing can be used to identify homophobic and transphobic texts and help make social media a safer place. In this paper, we explore the use of Support Vector Machine, Random Forest Classifier, and Bert Model for homophobia and transphobia detection. The best model was a combination of LaBSE and SVM, achieving a weighted F1 score of 0.95.

1 Introduction

The advent of social media platforms has revolutionized the way we interact, share, learn, express, and build our views and ideas. As these platforms have made communication with a large audience incredibly easy and available to everyone, free speech has become a governing concept of the virtual social realm. However, this newfound freedom has also posed its own threats. One major challenge of social media is hate speech. Hateful comments directed at minorities and vulnerable groups pose a significant threat because they can perpetuate existing prejudices and stereotypes, normalize or incite discrimination, and alienate these groups.

Homophobia and transphobia encompass a range of negative attitudes and feelings towards people based on their sexual orientation or gender identity. Homophobia refers to the fear, hatred, or prejudice against homosexuality, while transphobia involves discrimination against transgender individuals.

Research (Huebner et al., 2021) shows that sustained exposure to homophobic attitudes and behaviors can increase a person's stress levels. Studies (Wang et al., 2018) have demonstrated that adolescent victims of cyberbullying are more likely to experience depression and anxiety than adolescent non-victims. Disparities in mental health among LGBTQ youths persist into adulthood and adversely affect their development in social relationships, academic achievements, and self-concepts.

Sexual minorities also often make greater use of the internet as a result of seeking specific socialization environments in which they can meet other people with the same sexual orientation or avoid face-to-face social rejection and homophobic bullying (Gamez et al., 2021). This makes it even more necessary to address the problem of anti-LGBT hate speech.

The task given to us is detection of homophobia and transphobia in social media comments (Chakravarthi et al., 2023). In this paper, we have used the Language-Agnostic Sentence Embedder (LaBSE), along with Support Vector Machine (SVM). LaBSE encodes sentences in a way that captures their semantic meanings across multiple languages, enabling it to capture nuances and context more accurately than models that do not consider cross-lingual semantics. Along with this, SVM is used, known for its ability to handle high-dimensional data and effectively separate different classes. The combination of LaBSE and SVM constitutes an ensemble approach, where the strengths of both components are leveraged. Ensemble methods often result in better performance than individual models.

The paper is organized as follows: In Section 2, related works identified through a literature survey are presented. Section 3 offers an overview of the dataset, while Section 4 elaborates on the methodology employed for the task. The results are discussed in Section 5, and finally, Section 6 presents the concluding remarks.

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion, pages 266–271, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_041

2 Related Work

Debora Nozza (Nozza et al., 2022) proposed a solution for homophobia and transphobia detection based on data augmentation and ensemble modeling for high class imbalance dataset This task used large language models (BERT, RoBERTa, and HateBERT) and used the weighted majority vote on their prediction. This obtained 0.48 and 0.94 for macro and weighted F1-score, respectively.

Sushil Ugursandi and Anand Kumar M (Ugursandi and Anand Kumar, 2022) analyzed social media texts such as comments from YouTube to detect homophobic sentiments using deep learning or machine learning models. In this work, a 6-layer classification model was used and and an F1-Score of 0.5 on multi-class classification and 0.97 on homophobic/transphobic classification was achieved.

Konstantinos Perifanos and Dionysis Goutsos (Perifanos and Goutsos, 2021) employed transfer learning and fine-tuning of Bidirectional Encoder Representations from Transformers (BERT) and Residual Neural Networks (Resnet) for hate speech classification. They produced a high accuracy score of 0.970 and f1-score of 0.947 in racist and xenophobic speech detection.

Sunil Saumya and Ankit Kumar Mishra (Saumya and Mishra, 2021) analyzed social media texts, including comments from YouTube, to detect homophobic sentiments using deep learning and machine learning models. They employed a 6-layer classification model, achieving an F1-score of 0.5 for multi-class classification and 0.97 for homophobic/transphobic classification.

Shanita Biere (Biere et al., 2018) used a Convolutional Neural Network classifier to assign tweets to the categories : hate, offensive language, and neither. The model gave an accuracy of 0.91, precision of 0.91, recall of 0.90 and a F-measure of 0.90.

In another paper,(Lu et al., 2023) J. Lu and H. Lin utilized Dual Contrastive Learning to address the challenges posed by complex semantic information in hate speech and the imbalanced distribution between hate speech and non-hate speech data. The experimental results outperformed state-of-the-art models.

Orestes Appel (Appel et al., 2016) conducted sentiment analysis using a sentiment lexicon enhanced with the assistance of SentiWordNet, and fuzzy sets to estimate the semantic orientation polarity and its intensity for sentences. The results of the hybrid method was compared with Naïve Bayes and Maximum Entropy techniques. The hybrid method emerged to be the better performer. In addition, it is shown that when applied to datasets containing snippets, this method performs similarly to state of the art techniques.

Bharathi Raja Chakravarthi (Chakravarthi et al., 2022) addresses the challenge of limited resources for studying homophobia and transphobia detection. They propose a solution involving data augmentation through Pseudolabeling. This approach involves transliterating code-mixed text into the parent language, enhancing model performance on a newly generated dataset.

3 Dataset

The datasets provided were to us in Task A of Homophobia/Transphobia Detection in social media comments:-LT-EDI-2023 [1]. It consisted of social media comments in English where each comment had a label corresponding to it. The labels given were 'Non-anti LGBT+ content', 'Homophobia' and 'Transphobia'. The data was divided into three parts: training, development and testing, consisting of the columns 'text' and 'category'. The testing dataset was not provided with labels. The task was to predict the labels on our own. The training dataset consisted of 3164 entries with 2978 for Non-anti-LGBT+ content, 179 for Homophobia and 7 for Transphobia. The development dataset consisted of 792 entries with 748 for Non-anti-LGBT+ content, 42 for Homophobia and 2 for Transphobia. The train and dev datasets were used to train the model which was then tested on the test dataset. The test dataset had 991 entries. The development and training datasets given were severely imbalanced.

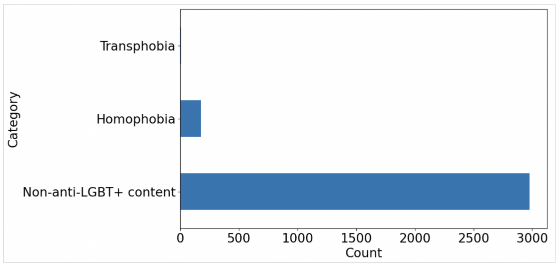

Figure 1: Train Dataset

[1] https://codalab.lisn.upsaclay.fr/competitions/11077

Label	Example	Instances in Train	Instances in Development
Non anti-LGBT+ content	Archana Shree what	2978	748
Homophobia	Shoot him all Dust bin	179	42
Transphobia	Hey seriously I thought She was a Transgender	7	2

Table 1: Dataset Description

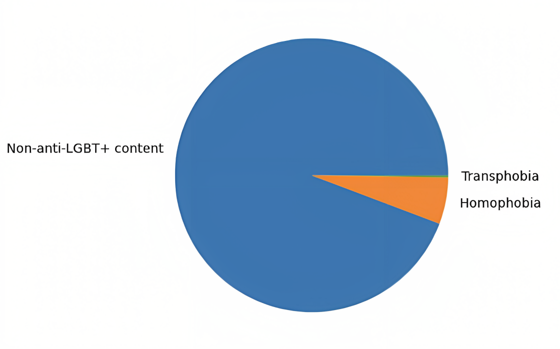

Figure 2: Dev Dataset

4 Methodology

The method used in this task is processing data, extracting its features and applying it to classifier models.

4.1 Data preprocessing

Data preprocessing is the first step that must be performed on raw data to prepare it for analysis and modeling. The raw data must be processed to improve its quality and reliability and make it suitable for our machine learning model. First the data must be cleansed. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. These errors are called as noise and can diminish the performance of our machine learning models. Any data must be cleansed before we begin our work with it. The models used in this task are based on finding common words that occur in the texts corresponding to each label. So the next step is manipulating the data to be suitable for the models. The following procedure was adopted in the task:

1) Checking for null values: First the data is checked for missing values. Most machine or deep learning models require you to clean up the data of null values before it is used. If missing values are present, they are dropped.

2) Removing punctuation and special characters: Since the models used focus on finding common words, punctuations and special characters are meaningless and are considered as noise in the data. A list of punctuations from the string library are used to remove the punctuations and special characters from the text.

3) Converting to lowercase: The text is converted to lowercase to standardize the text data so that different forms of the same words are considered the same(For example, “Eating”, “EATING”, “eating). By converting to lowercase, the analysis becomes case sensitive and enables an accurate frequency counts of words.

4)Removing stop words: Stop words are filler words that are insignificant and do not carry any meaning in the context of the task (for example: the, a, are, in). Stop words occur so frequently that they can skew the result of the models. A list of predefined stop words is used to remove the stop words from the text.

4.2 Embeddings and Feature Extraction

Word embeddings are representations of words as vectors in vector space such that words with similar meanings are closer together. These embeddings can be used for a wide range of NLP tasks, such as text classification, semantic similarity, clustering, and information retrieval.

Features are the individual measurable properties of the data that are used as input variables for a model. Features provide information or attributes that help the model understand and make predictions or classifications based on the patterns and relationships present in the data. The selection and quality of features play a crucial role in the performance and accuracy of the machine learning model.

Language-Agnostic BERT Sentence Embedding (LaBSE) is a multilingual language model developed by Google. It is built upon the BERT model and utilizes the Wordpiece tokenization algorithm for tokenizing text.

LaBSE follows a dual encoder architecture, meaning it has two separate encoders. These en-

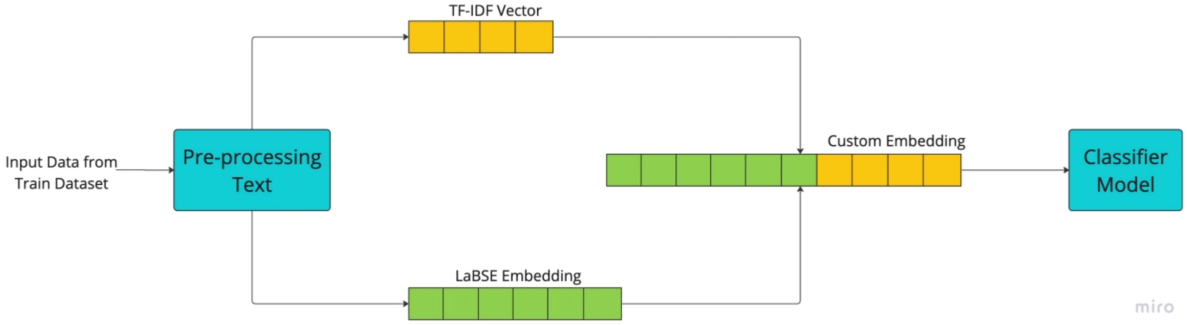

Figure 3: Methodology

coders independently process source and target sentences. The encoded representations of the sentences are then passed through a scoring function that ranks them based on their similarity. This technique enables LaBSE to store similar sentences close to each other in a shared embedding space.

LaBSE learns to capture universal semantic patterns across different languages. This enables the model to generate meaningful sentence embeddings for sentences in multiple languages, even for languages that were not included in the training data.

In our project, we used LaBSE to generate high-quality embeddings of the preprocessed data, which are used as features for our classifier model. The classifier model classifies the given text into its corresponding labels.

4.3 Models Used

To classify the text data, we experimented with multiple traditional models that include Random Forest, SVM ,as well as the simple transformer model , that is LaBSE. After evaluating the metrics of multiple models, we focused on combining the LaBSE feature extraction model along with the SVM classifier.

4.3.1 Random Forest

Random forest is a supervised machine learning algorithm used for classification that is based on the concept of ensemble learning. Ensemble learning is the process of combining multiple classifiers to solve a complex problem and improve the performance of the model. RF contains a number of decision trees on various subsets of the dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree. based on the majority votes of predictions, it predicts the final output.

4.3.2 Support Vector Machine

Support vector machine or SVM is a supervised machine learning model that is popularly used for classification. SVM algorithm finds an optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space, such that the margin between points of different classes is maximized. The accuracy of an SVM classifier model can be increased by increasing the number of dimensions.

5 Result and Analysis

After evaluating the metrics of multiple models, we focused on combining the LaBSE feature extraction model along with the SVM classifier as this gave us the best results.

5.1 Performance Metrics

Three performance metrics were used for evaluating the task, namely Recall, Precision and F1 score. The macro average and the weighted average of these metrics were also calculated.

Precision: It is the ratio of correctly classified data points to the total number of data points that have been predicted to be of that class. High precision indicates that the model makes fewer false positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall: It is the ratio of correctly predicted positive instances out of all actual positive instances. High recall indicates that the model successfully identifies a larger portion of the actual positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced evaluation of the model by considering both precision and recall. It balances precision and recall, making it useful when both measures are important, such as in imbalanced datasets, or when avoiding false positives and false negatives is crucial.

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (3)$$

The models were evaluated on development dataset. Random Forest yielded a weighted F1 score of 0.92. Support Vector Machine yielded a weighted F1 score of 0.93. However, the best results were shown when LaBSE feature extraction was combined with SVM Classifier, yielding an F1 score of 0.95. This could be attributed to the advantage of using ensemble techniques.

	Precision	Recall	F1-Score	Support
0	0.95	1.00	0.97	748
1	0.67	0.05	0.09	42
2	0.00	0.00	0.00	2
Accuracy			0.95	792
Macro Avg	0.54	0.35	0.35	792
Weighted Avg	0.93	0.95	0.92	792

Table 2: RF Classification Report

	Precision	Recall	F1-Score	Support
0	0.95	0.99	0.97	748
1	0.44	0.10	0.16	42
2	0.00	0.00	0.00	2
Accuracy			0.94	792
Macro Avg	0.46	0.36	0.38	792
Weighted Avg	0.92	0.94	0.93	792

Table 3: SVM Classification Report

	Precision	Recall	F1-Score	Support
0	0.96	0.98	0.97	748
1	0.70	0.45	0.55	42
2	0.00	0.00	0.00	2
Accuracy			0.96	792
Macro Avg	0.55	0.48	0.51	792
Weighted Avg	0.94	0.96	0.95	792

Table 4: SVM with LaBSE Classification Report

In the classified test data submitted for Task A of Homophobia/Transphobia Detection in social media comments:-LT-EDI@RANLP-2023 on English Dataset, SVM with LaBSE yielded an F1 score of 0.95.

Our submission was ranked the 5th place in Task A of Homophobia/Transphobia Detection in social media comments:-LT-EDI@RANLP-2023 on English Dataset.

6 Conclusion

In this paper, we have presented traditional classification models coupled with LaBSE, a pre-trained language agnostic BERT model, for the classification of Non-Anti LGBT, Homophobia and Transphobia comments on the data given by Dravidian-LangTech in the English language. The traditional models explored were Random Forest and Support Vector Machine and it was found that SVM yields a higher F1 score of 0.95. We believe that we can improve the accuracy of our results using more sophisticated models such as deep learning architectures (e.g., convolutional neural networks, recurrent neural networks, transformers) and combining predictions from multiple models or model variations through techniques like ensemble learning.

References

Orestes Appel, Francisco Chiclana, Jenny Carter, and Hamido Fujita. 2016. A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124.

Shanita Biere, Sandjai Bhulai, and Master Business Analytics. 2018. Hate speech detection using natural language processing techniques. *Master Business AnalyticsDepartment of Mathematics Faculty of Science.*

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. *International Journal of Information Management Data Insights*, 2(2):100119.

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.

Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and

Bo Xu. 2023. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2787–2795.

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.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Sunil Saumya and Ankit Kumar Mishra. 2021. Iiit_dwd@ lt-edi-eacl2021: hope speech detection in youtube multilingual comments. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 107–113.

Sushil Ugursandi and M Anand Kumar. 2022. Sentiment analysis and homophobia detection of youtube comments. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.