Leveraging Transfer Learning Techniques for Homophobia and Transphobia Detection

MSc Research Project
Data Analytics

Syed Ebrahim Abdul Kareem
Student ID: x20232616

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Syed Ebrahim Abdul Kareem
Student ID:	x20232616
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	01/02/2023
Project Title:	Leveraging Transfer Learning Techniques for Homophobia and Transphobia Detection
Word Count:	5037
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Syed Ebrahim Abdul Kareem
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Leveraging Transfer Learning Techniques for Homophobia and Transphobia Detection

Syed Ebrahim Abdul Kareem
x20232616

Abstract

The proliferation of hate speech and abusive content on social media makes it unsafe place for networking. Unlike past, discussion on LGBTQ+ topic is openly happening on public forums at present. People have started to openly declare their sexual orientation and gender identification. Many nations have legalized such acts which were once considered a crime or sin. Due to anonymity and aversion on LGBTQ+ community, some users spread hatred towards them on social media. To ensure equality, inclusiveness and diversity, social media companies must detect and moderate such posts. This will ensure harmony and maintain decorum on their platforms. In Natural Language Processing domain, one of the most active areas of research is text classification. Over the last few years, Transformer based models are widely used for this task as they provide better results than conventional Machine Learning models. In this research, RoBERTa, DistilBERT and mBERT were implemented and their performance is compared with past works. The results shows that DistilBERT outperformed RoBERTa with a macro avg F1 score of 0.47 on English dataset. This is slightly better than conventional ML models like Logistic Regression and SVM. On the Tamil dataset, mBERT model achieved a macro avg F1 score of 0.83 which is 3.75% improvement over previous studies.

1 Introduction

1.1 Background & Motivation

The development of information technology and widespread access to internet has resulted in the world becoming more and more interconnected. This is happening through social media platforms like Facebook, TikTok, Instagram, YouTube, WhatsApp and Twitter. These platforms provide anonymity to users who share and post contents. These contents include videos, images, audio files and text data. Huge volume of data is generated every day from these websites. The variety and huge volume of data generated can be used by Natural Language Processing researchers to solve problems like understanding user behavior, their sentiments on certain topics or issues. While the main idea behind social media is to provide a platform for common people to engage with others and express their views on anything, some users are tempted to share fake news or abusive contents. This act affects the harmony and decorum of the platform. It makes the platform unpleasant and unsafe for networking. Most of the NLP works done in the past have focused on detecting the fake news or hate speech in general, there is little or no research that were carried out to detect the hate speech and abusive content spread against the LGBTQ+ community. LGBTQ+ is a term used to collectively denote those people who identify themselves as one of the following – Lesbian, Gay, Bisexual, Transgender and Queer. In early days, discussion on sexual orientation and gender identification was considered as taboo on public forum. In modern culture, discussion on these topics has increased. People use social media as a tool to discuss and declare their identification. Due to anonymity and various other reasons, some users use hatred, abusive and foul languages against this vulnerable community. This research tries to explore the use of deep learning techniques to detect homophobia and transphobia on social media. Homophobia refers to hatred, abuse, prejudice spread against homosexual persons or persons feeling attraction towards others of same gender. Transphobia refers to the hatred and abuse directed towards transgender people. (Damas and Mochetti; 2019) has discussed about vandalism on Wikipedia pages due to homophobia. Two groups of famous computer scientists were chosen. One group with scientists who identified themselves as LGBTQ openly and others who did not identify with LGBTQ. One example from each group - Alan Turing was in first group and Bill Gates was in second group. It was observed that vandalism on Wikipedia page of Alan Turing was more related to LGBTQ by trolling his sexual orientation while vandalism on Wikipedia page of Bill Gates was mostly related to blank pages and silly edits. This shows the intolerance of online users against this community.

In recent years, Natural Language Processing has been used on a variety of applications like Google search engine, Amazon Alexa and Apple Siri to process and understand human language. Deep learning techniques are producing better results when compared to the conventional machine learning algorithms. Google developed BERT model based on Transformer architecture to solve NLP tasks. This model was pre-trained on English language words from Book corpus and Wikipedia pages. Today, most of the English search query on google search engine uses BERT. Much research has been done to improve BERT and apply it to other languages. As a result there are many versions of BERT such as Robustly Optimized BERT (RoBERTa), Distilled version of BERT (DistilBERT), Multilingual BERT (mBERT) which was trained on 104 languages using Wikipedia pages, a light version of BERT (ALBERT) and Generalized Autoregressive Pretraining for Language Understanding (XLNET). These pretrained models were extensively used in NLP use cases like fake news detection, hate speech detection, etc. However, these models are yet to be used for homophobia and transphobia detection.

1.2 Research Question & Objective

1.2.1 Research Question

To what extent Transformer based pre-trained transfer learning models can effectively detect homophobia and transphobia on social media comments?

1.2.2 Research Objective & Contribution

The following objectives were set to deal the research question posted above:
1. A critical analysis on research and studies performed for homophobia and transphobia detection in the past.
2. Pre-processing the data to make it compatible for the pre-trained models: DistilBERT, RoBERTa and mBERT.
3. Implementing RoBERTa, DistilBERT and mBERT models for homophobia and transphobia classification task.
4. Performance evaluation of the implemented models.
5. Comparison of the implemented models using the evaluation metric - f1 score.
6. Comparison of the implemented models with the base line models.

1.3 Structure of the Report

This report is organized as follows: The second section discusses about previous studies and research performed on homophobia and transphobia detection, and similar Natural Language Processing tasks. The third section explains about the methodology used in this research. The fourth section explains the framework used and the stages involved. The fifth section deals with the details on implemented models. The next section deals with the evaluation of the models and results. The last section deals with conclusion and future work.

2 Related Work

The issue of homophobia and transphobia is prevalent on social media only in the recent times. Only limited number of works have been done on this topic in the past. (Chakravarthi et al.; 2021) were one of the first to create a dataset to promote and enable research on this topic. While most of the Natural Language Processing tasks used conventional machine learning models until few years backs, Deep Neural Network models have become popular in recent times because of its performance. This section discusses about the homophobia and transphobia detection, Machine Learning and Deep learning techniques used in NLP domain and Pre-trained models used by researchers for text classification tasks.

2.1 Homophobia and Transphobia detection

The abusive language used against LGBTQ+ community is an important social problem that needs to be addressed. However, very few works have been done in the past on this topic. This has led to limited availability of datasets on this topic. The authors of (Chakravarthi et al.; 2021) have made an effort to create a corpus of datasets that can be utilized to perform sentiment analysis on homophobia and transphobia. There are three datasets with different language settings. It contains user posted comments from YouTube videos that talks about LGBTQ+ community. The three settings include English, Tamil and Code-mixed (Tamil-English) datasets. This research paper focusses only on English and Tamil settings for sentiment analysis. The authors have build base line models using SVM, Random Forest classifier, Logistic regression, and Decision tree classifier. They have also examined the performance of different word embedding techniques like Term Frequency-Inverse Document Frequency (TF-IDF), Count Vectorizer, FastText, etc. In addition to this, the authors have also implemented Deep Neural Networks (DNN) like BiLSTM. Since, the dataset is highly imbalanced, macro avg F1 score was chosen as the evaluation metrics. Most of the models produced macro average score of less than 0.4. This indicates that there is scope for further improvement and class imbalance can be handled more effectively.

A model based on lexicon was suggested by (Anand et al.; 2019) to examine how a homosexual person's statement on their sexual orientation is perceived by friends and families on social media. For this study, tweets posted by homosexual persons that declared their sexual orientation were collected. For each tweet, a score is assigned based on the Wilson lexicon's polarity score for each word and the distance between the subject and that word. A major limitation of this approach is that those words that are not part of the lexicon will not contribute to the analysis. Hence, this approach is prone to information loss and not suitable for homophobia and transphobia classification.

2.2 Text classification using ML and DL methods

Analysis on sentiments expressed by social media users is continuously carried out in recent times. (Mittal and Patidar; 2019) have discussed different machine learning techniques that can be used to classify tweets on twitter. Techniques includes Naïve bayes, SVM, Maximum entropy and lexicon-based approach. Supervised techniques have a disadvantage that only sub-optimal results are achieved with insufficient data. Also, for models built using lexicon, the performance will be degraded if the texts in the samples that are part of the dataset are not part of the dictionary used. (Mantoro et al.; 2021) have developed a machine learning model to classify the tweets related to Papua movement. Naive bayes multinomial(NBM) algorithm was used to classify the tweets. The model achieved an overall accuracy of 94%.

The authors of (Ilmania et al.; 2018) have carried out aspect detection and sentiment analysis. Normally, text classification can handle sentences with only one polarity. Real time data may have sentences with more than one polarity. Aspect based sentiment classification can be used in such scenario. For aspect detection, a comparison was made between two models – bag of words with fully connected neural network and word embedding vector with Gated Recurrent Unit (GRU). For sentiment classification, lexicon-based feature extraction with Bi-GRU model was compared with word embedding vector and

CNN model. It was observed that the performance of GRU based models were better when compared to other models in both cases.

Social media is filled with lot of fake news that are spread by both individuals and institutions. They spread faster than truth and affects the society. A system has been proposed by (Bhutani et al.; 2019) to detect fake news. Word embedding techniques like Term Frequency Inverse Document Frequency(TFIDF), count vectorizer and TFIDF combined with cosine similarity were used for feature extraction. The authors have chosen Random Forest and Naive bayes models for implementation. A combination of TFIDF with cosine similarity was more efficient in extracting features from text data when compared to other word embedding techniques. It performed well on both the algorithms.

A sentiment classification model based on deep learning algorithms to classify reviews posted on social media was discussed by (Cheng and Tsai; 2019). First, the authors have performed data pre-processing by splitting sentences into words, slang correction, removing emoji and emoticon. Then, tokenization was carried out with GloVe and word2vec. Then, three Deep Learning algorithms namely LSTM, BiLSTM and Bi-GRU were used for implementation and performance were compared based on recall, precision, accuracy and F1 score. BiLSTM model outperformed all the other models and achieved F1-score of 87.29%. The authors of (AlSalman; 2020) have discussed about building a machine learning model to classify tweets made on Arabic language. Data pre-processing was performed using N-grams and Khoja stemmer. Khoja stemmer performs stemming operation specifically for Arabic language. Vectorization was carried out using TF-IDF. Authors have used Discriminative Multinomial Naive Bayes(DMNB) algorithm for implementation. This model provided an improved accuracy of 87.5% when compared to previous models.

Due to limited availability of lexicons for under resourced languages like Hindi, lexicon-based sentiment analysis is cumbersome for such languages. Under resourced languages can leverage the application of Machine Learning and Deep Learning concepts in NLP tasks like sentiment analysis. (Goel and Batra; 2020) have developed a deep learning model to perform sentiment classification on Hindi tweets. Word embedding techniques like word2vec, doc2vec and TF-IDF were used. Three classifiers were built using SVM, Logistic regression and LSTM. In this, LSTM performed better with an accuracy of 65% to classify Hindi tweets into positive and negative.

Sentiment analysis can be used in e-commerce business to understand how customers are perceiving the company's products by analyzing online user reviews. (Vimali and Murugan; 2021) have proposed a sentiment analysis model to classify user reviews into positive and negative class. The model is trained with reviews scrapped from Amazon website. Word2vec along with Bi-LSTM was used to build the model. It provided an accuracy of 90.46%. While most of the use cases deal with monolingual data for text classification, there are few instances where the text data generated is code-mixed. (Yadav et al.; 2020) discusses an approach to perform sentiment analysis on Hindi-English code-mixed data. Two approaches were discussed by the authors. In first approach, an ensemble of SVM, Linear regression, Naïve Bayes and Stochastic Gradient Descent classifiers. Bi-LSTM was used in the second approach. Results indicated that ensemble technique (accuracy of 74%) performs slightly better than Bi-LSTM (accuracy of 73%)

2.3 Pre-trained Transfer learning models

A comparison on the performance of different versions of pretrained Transformer models were made by (Adoma et al.; 2020). This study compares Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa), Generalized Autoregressive Language Model (XLNET) and Distilled version of Bidirectional Encoder Representations from Transformers (DistilBERT). These models were implemented to classify texts on ISEAR datset into joy, guilt, shame, anger, fear, sadness and disgust. It was found that RoBERTa model performed well in comparison to other models. The accuracy was 74.31%, 72.99%, 70.09% and 66.94% for RoBERTa, XLNET, BERT and DistilBERT models respectively. This study also found that XLNET model required more computation power and training time while the DistilBERT model required very less training time with less computation power.

The authors of (Ghanghor et al.; 2021) have discussed an approach for hope speech detection on YouTube comments. In this study, experiments were carried out on three datasets – Tamil, English and Malayalam. Pretrained Transformer models like BERT, Multilingual BERT (mBERT), XLMR (XLM-RoBERTa) and IndicBERT were implemented. Among these mBERT model outperformed other models with an F1 score of 0.60 and 0.83 for Tamil and Malayalam dataset respectively. Both mBERT and BERT models performed almost similarly on English dataset with F1 score of 93%

A comparison of various transfer learning models was carried out by (Rajapaksha et al.; 2021) on detecting clickbaits on social media. Clickbaits refers to showing eye catchy headlines to generate more page views but with deceptive news content. The authors have compared the performance of RoBERTa, BERT and XLNET models to classify news clickbaits. For this study, Webis Clickbait dataset was used. Various modifications and novel approach were done to the base models for fine tuning. Among these, RoBERTa outperformed all the other models and obtained 19.12% more accuracy than baseline model.

3 Methodology

This research project focusses on building a classifier to detect homophobia and transphobia on social media. Cross-Industry Standard Process for Data Mining (CRISP-DM) and Knowledge Database Discovery (KDD) are the two widely used methodologies for data mining. This project uses KDD methodology. The KDD methodology consists of six stages which are structured and iterative in nature. Moving between back and forth the stages might be required. The steps involved in KDD can be seen in Figure 1.

3.1 Goal Setting and Business Understanding

Understanding the business requirements is very important to come up with the right solution for any problem. Hence, this is considered as the first and foremost step in KDD methodology. First the problem should be understood well and then converted into an analytical problem statement. Social media users use anonymity to spread abusive and hatred content hiding behind the screen. These are targeted towards vulnerable sections

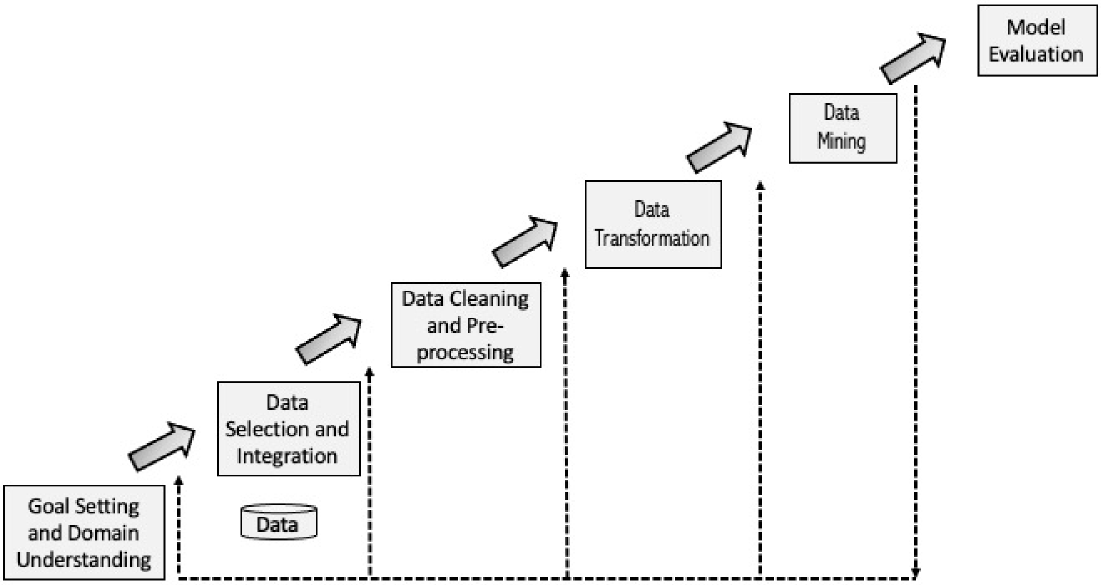
Goal Setting and Domain Understanding
Data
Data Selection and Integration
Data Cleaning and Pre-processing
Data Transformation
Data Mining
Model Evaluation

Figure 1: KDD methodology for Data Mining projects

like LGBTQ+ community. This spoil the decorum and harmony of the social media platform. The social media companies find it difficult to moderate such contents. It is impossible to moderate such contents manually as it takes a lot of time and manpower. Hence, there should be a Machine Learning based automated system in place which can detect homophobia and transphobia for content moderation.

3.2 Data Selection and Integration

This stage of the KDD methodology deals with the collection and understanding of the dataset. This research project uses the dataset created by [1] to build a classifier to detect homophobic and transphobic comments. There are three datasets compiled under three different settings – English, Tamil and English-Tamil code mixed. This project will make use of two datasets – English and Tamil. The dataset contains user comments extracted from YouTube videos that talks about topics related to LGBTQ+ and their class labels. The comments are annotated as one of the following – Non-Anti-LGBTQ+ content, Homophobic and Transphobic. There are 4946 comments for English dataset and 3977 comments for Tamil dataset with 3 classes in both the datasets. Figure 2 and Figure 3 shows the class-wise distribution of dataset for English and Tamil dataset respectively. It can be noticed that both the datasets are highly imbalanced and representation of Transphobic class in English dataset is negligible.

3.3 Data Cleaning and Pre-processing

In Natural Language Processing, data cleaning and pre-processing is an important step where all the noises and irrelevant information from the data are removed before feeding it into the machine learning model. Usually text data contains spelling mistakes, stop words, punctuations, line breaks, recurring words, emojis, emotions and informal words

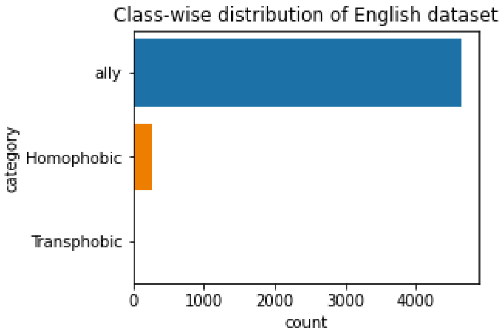

Figure 2: Class-wise distribution of English dataset

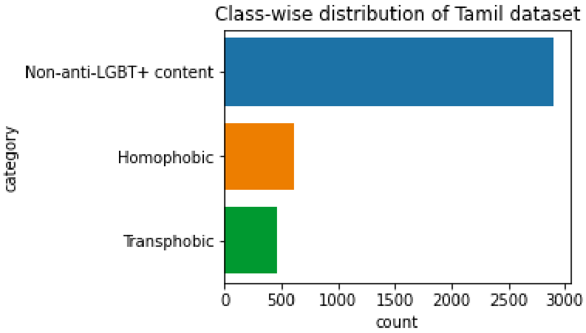

Figure 3: Class-wise distribution of Tamil dataset

like “ROFL”, “ASAP”, “LOL”, “FYI”, etc. The text clean-up can be performed by using pre-built python libraries. After cleaning up, the text is converted into sentences. This process is called tokenization. Next, word tokenization is performed by splitting sentences into words. Finally, all the inflected words are converted to their base words. For example, the word “came” is converted into its base word “come”. This process is called lemmatization. These data preparation steps should be followed before performing data transformation.

3.4 Data Transformation

In this step, the text data is converted into machine readable format. Since the machine learning algorithms cannot process text data directly, it needs to be converted into numbers. This process is called vectorization or word embedding. It is in this stage; the feature extraction is done. All the information from the text data is extracted and stored as numbers. There are many tokenizers available in python which does this job. This research project uses RoBERTa and DistilBERT tokenizer for English dataset and BERT tokenizer for Tamil dataset. The output of the tokenizer will be tensors or arrays that are compatible with the specific pre-trained language models.

3.5 Modeling

In this stage of the KDD methodology, the right model is selected based on literature review and adopted in the project. Based on the literature review conducted, many machine learning and deep learning techniques were analysed in depth and finally decided to leverage the potential of Transformers based pre-trained models from BERT family as their performance were better than conventional ML models for NLP tasks and less explored for homophobia and transphobia detection. RoBERTa and DistilBERT will be used for classifying English dataset and Multilingual BERT (mBERT) will be used for Tamil dataset. Simple Transformers and TensorFlow frameworks will be used for implementing these models.

3.6 Model Evaluation

This is the last stage in KDD methodology, in which the relevant evaluation metrics is selected for evaluating the performance of models. Since the datasets are highly imbalanced, we will use macro average F1-score (Melton et al.; 2020) to evaluate the models. F1-score is defined as the harmonic mean of recall and precision. F1-score ensures that the model performs well in terms of both precision and recall. Macro average F1-score is defined as the arithmetic mean of F1-score of each class. The macro average F1-score computation gives equal weightage to all the classes irrespective of their occurrences.

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1}$$

4 Design Specification

This project will use a three-tier framework for homophobia and transphobia detection using Transformers based pre-trained transfer learning models as shown in Figure 4. The framework consists of three layers – data layer, business logic layer and final evaluation layer. In the data layer, the data is loaded into Google Colab Pro and Kaggle Notebook for English dataset and Tamil dataset respectively. Both uses Graphics Processing Units (GPU) for cleaning the data, pre-processing and tokenization. The business logic layer covers the model building part where different models were trained. The final evaluation layer covers the model evaluation. Here, appropriate evaluation metrics is selected for evaluation of models and compared.

5 Implementation

This section explains about the implementation of the Transformers based pre-trained transfer learning models to classify social media comments based on their sentiment to LGBTQ+ community. All the steps involved in building the classifier is discussed in detail from data pre-processing to model building. Furthermore, the setup configuration used for implementation of the models is given in Table 1.

5.1 Data Pre-processing

In NLP, data cleaning and pre-processing is very crucial where irrelevant information is discarded and valuable information is retained. The following data pre-processing steps were followed for English dataset.

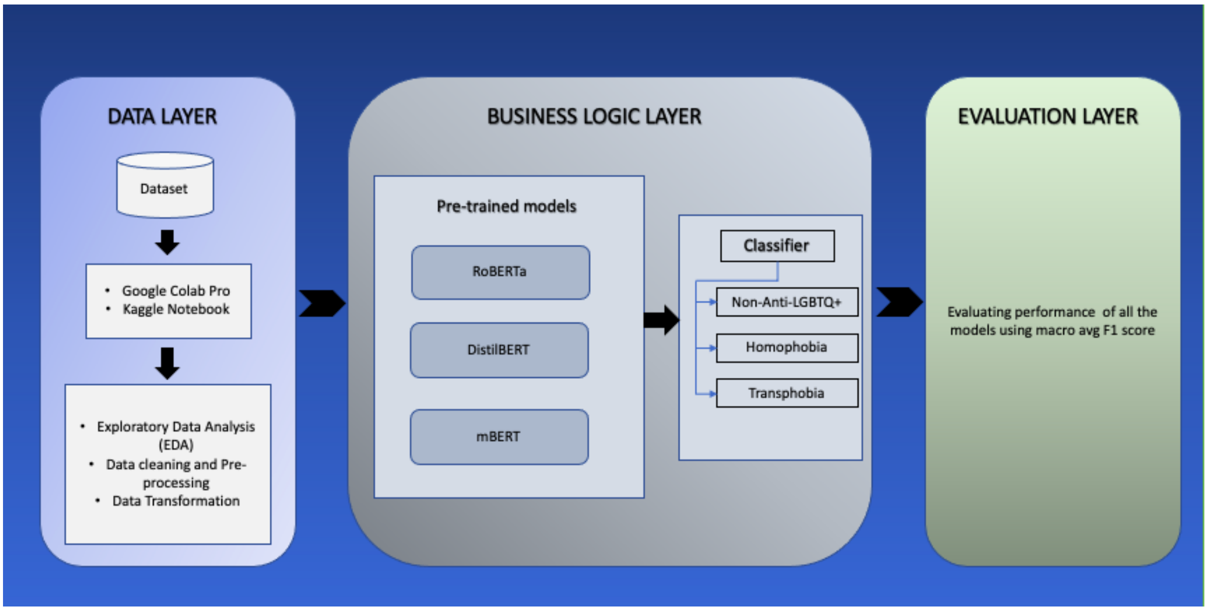

Figure 4: Design framework used in this research project

Table 1: Setup Configuration

IDE	Google Colab Pro and Kaggle Notebook
Computation	GPU
Type	Tesla P100-PCIE-16GB
Number of GPU	1
Programming language	Python
Modules	Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, Transformers, NLTK, Emoji, Num2words and Simpletransformers
Framework	Tensorflow

5.1.1 Punctuation removal

The text data may contain punctuations which doesn't add any value to the context or difficult to process. Hence, punctuations are removed using tokenizer function from Keras pre-processing module.

5.1.2 Emoji conversion

Emoji's are used to express one's thoughts or feelings in a pictorial representation. Most of the interaction happening online use emojis. It carries meaning to the context in which it is used. For example, a person feeling loved can be conveyed using a smiling face with hearts emoji 😍. We use emoji library in python to convert emoji into description in English words. An example is given below.
Before – Respect to all 🙏 🙏 🙏
After - Respect to all **folded_hands folded_hands folded_hands**

5.1.3 Stop word removal

Stop words are words that are commonly used in human language. They can be pronoun, articles, conjunctions, etc. These words should be removed from the input data as these doesn't add any meaningful information to the modeling process. For example: "an", "the", "and" are stop words in English language. This step is performed with the help of English stop words corpus from NLTK library.

5.1.4 Number to words

In this step the numbers present in the texts are converts to words. This is performed using num2words module in python. An example is given below.
Before – I ate **2** apples
After – I ate **two** apples

5.1.5 Lemmatization

Lemmatization is a normalization technique applied on text data in Natural Language Processing tasks. In this technique, inflected words are converted to their base root words without changing the meaning. This is performed with the help of NLTK library. An example is given below.
Before – I **ate** two apples
After – I **eat** two apples

5.2 Implemented Models

Based on the literature review, Transformers based pre-trained transfer learning models provide better results on text classification and other NLP problems. This approach is used extensively in the recent times. In this research project, fine-tuned RoBERTa and DistilBERT model on English dataset and mBERT model on Tamil dataset were implemented.

5.2.1 RoBERTa

Robustly Optimized BERT approach (RoBERTa) is a modified version of BERT in which the next sentence prediction objective was removed. This makes it easier to train in terms of computation time and get rid of the NSP loss present in BERT. It was developed by AI team (Liu et al.; 2019) at Facebook in 2019. It was trained on a larger dataset with longer sequences of English texts when compared to BERT. It inherits language masking strategy from BERT. Byte Pair Encoding (BPE) technique is used for tokenization. In this research paper, a fine-tuned RoBERTa base version is used. Since the dataset is imbalanced, the number of instances of minority classes were increased using Easy Data Augmentation technique introduced by (Wei and Zou; 2019). A dense layer was introduced with 64 neurons and the output layer has 3 neurons with "softmax" activation function. Two drop out layers were introduced, one between pre-trained model output and dense layer and another between the dense layer and output layer for generalization of the model. Dropout rate was set at 0.4 and 0.2 respectively. The model is trained for 5 epochs with Adam optimizer of learning rate 2e-5 and "CategoricalCrossentropy" as the loss function.

5.2.2 DistilBERT

A distilled version of BERT approach is known as DistilBERT. It was first introduced by (Sanh et al.; 2019). It is a lighter, smaller and faster version built by distilling the base BERT model. When compared to BERT base model, it uses 40% less training parameters, trains 60% quicker while retaining 95% of BERT base model's performance. The token type embeddings and pooler are not used in DistilBERT which were originally used in BERT model. Also, 50% of the layers that BERT uses are retained by DistilBERT approach. The limitations of BERT model like fixed input length and word piece embedding are mitigated in DistilBERT version. A triple loss system has been adopted which includes student loss, cosine loss and masked language modeling loss. This project adopts DistilBERT as one of the models for implementation for its low resource requirement and performance on text classification tasks. Downsampling and Upsampling were performed on the train data to mitigate class imbalance. Maximum length of input vector was set as 241 and those sentences that had length less than 241 were padded. The model is fine tuned by adding a Bi-LSTM layer and dense layer in between the output of pre-trained model and the final output layer. With ”Sparse categorical cross entropy” as the loss function and Adam optimizer, the model was trained for 5 epochs.

5.2.3 Multilingual BERT (mBERT)

BERT multilingual base model is a multilingual version of BERT base (Devlin et al.; 2018) that can handle NLP problems for English as well as other languages. It was trained on more than 100 different languages using Wikipedia texts. It follows the masked language modeling objective (MLM). In this research, simple transformers library which was built on top of HuggingFace was used for implementing the model. Maximum sequence length is initialized as 180 and learning rate as 2e-5. The model was trained for 5 epochs.

6 Evaluation & Results

The evaluation metrics selected in this research project for evaluating the performance of different pre-trained Transformer models is F1-score. Since, the dataset is highly imbalanced, instead of just relying on overall accuracy, this research uses macro average F1-score for evaluating model performance. Computationally, macro avg F1-score gives equal weightage all three classes – Non-Anti-LGBTQ+ content, Homophobia and Transphobia irrespective of their frequency in the dataset.

6.1 English dataset

Table 2 shows the results achieved by RoBERTa model on English dataset. The macro F1-score achieved by RoBERTa model is 0.33. Overall accuracy stands at 0.92

Table 3 shows the performance of DistilBERT model on English dataset. In this model, upsampling and downsampling were performed to balance the classes. The overall accuracy and macro F1 score for DistilBERT model is 0.93 and 0.47 respectively.

Table 4 shows the results achieved by base line models created by (Chakravarthi et al.; 2021). It can be noticed that both the Logistic Regression and Support Vector Machine

Table 2: Evalution metrics for RoBERTa model on English dataset

Evaluation metrics	**Value**
Accuracy	0.92
Macro Avg Precision	0.33
Macro Avg Recall	0.33
Macro Avg F1 score	0.33

Table 3: Evalution metrics for DistilBERT model on English dataset

Evaluation metrics	**Value**
Accuracy	0.93
Macro Avg Precision	0.47
Macro Avg Recall	0.46
Macro Avg F1 score	0.47

(SVM) models have provided the best results on English dataset with a macro F1-score of 0.46

Table 4: Results achieved by baseline models on English dataset

Model	**Accuracy**	**Macro avg F1 score**
Logistic Regression	0.90	0.46
Naïve Bayes	0.75	0.39
Random Forest	0.94	0.44
SVM	0.91	0.46
Decision Tree	0.93	0.38
Bi-LSTM	0.94	0.32
mBERT	0.06	0.04

6.2 Tamil dataset

Table 5 shows the performance of mBERT model on Tamil dataset. From the table, it can be inferred that macro F1-score and overall accuracy for mBERT model was 0.83 and 0.92 respectively.

Table 5: Evalution metrics for mBERT model on Tamil dataset

Evaluation metrics	**Value**
Accuracy	0.92
Macro Avg Precision	0.82
Macro Avg Recall	0.83
Macro Avg F1 score	0.83

Table 6 shows the performance of different models on Tamil dataset implemented by (Chakravarthi et al.; 2021). It can be noticed that both Random Forest and Support Vector Machine (SVM) model achieved a macro F1-score of 0.80.

Table 6: Results achieved by baseline models on Tamil dataset

Model	**Accuracy**	**Macro avg F1 score**
Logistic Regression	0.84	0.64
Naïve Bayes	0.72	0.56
Random Forest	0.92	0.80
SVM	0.89	0.80
Decision Tree	0.78	0.47
Bi-LSTM	0.89	0.31
mBERT	0.16	0.28

6.3 Discussion

The objective of this research is to leverage the potential of pre-trained Transformer models to perform sentiment analysis on YouTube comments by classifying them into Non-Anti-LGBTQ+ content, Homophobia and Transphobia. (Chakravarthi et al.; 2021) have used the same dataset and worked on conventional Machine Learning models like Logistic regression, Naïve Bayes, Random Forest classifier, Support Vector Machine (SVM), Decision Tree classifier and Deep Learning models like Bidirectional LSTM and Multilingual BERT model. Table 7 shows the comparison of F1-score achieved by Transformer models and conventional Machine Learning models on English dataset. It can be inferred that DistilBERT model has achieved the highest Macro avg F1-score of 0.47 which is closer to 0.46 achieved by LR and SVM classifiers. However, the performance of RoBERTa model is not satisfying. The implementation of Easy Data Augmentation technique has not created any significant effect on the model's performance. This can be due to the fact that EDA is not suitable for pre-trained models since they are already trained on huge amount of data as per (Wei and Zou; 2019). Table 8 shows the comparison of F1-score obtained by mBERT and conventional ML models for Tamil dataset. It can be observed that mBERT model has outperformed Random Forest and Support Vector Machine (SVM) models.

Table 7: Comparison of models for English dataset

Model	**Macro avg F1 score**
Logistic Regression	0.46
SVM	0.46
RoBERTa	0.33
DistilBERT	**0.47**

Table 8: Comparison of models for Tamil dataset

Model	**Macro avg F1 score**
Random Forest	0.80
SVM	0.80
mBERT	**0.83**

7 Conclusion and Future Work

The hate speech or abusive posts against vulnerable communities like LGBTQ+ on social media platforms is increasing day by day as many people are coming forward and self-declaring their identity. Hence, it is crucial for the social media companies who own these platforms to ensure inclusiveness, diversity and equality by moderating contents that target specific set of people.The main motive of this research project is to explore the potential of Transformer based pre-trained transfer learning models to detect homophobia and transphobia on social media comments. RoBERTa and DistilBERT models were created to classify comments in English. In this study, DistilBERT outperformed RoBERTa and other conventional Machine Learning models that were built in the past with a macro avg F1-score of 0.47. Multilingual BERT model was implemented on Tamil dataset. It provided a macro F1-score of 0.83 which is better than the conventional Machine Learning models that were built in the past. In current cosmopolitan society, code mixed texts are more common as people use multiple languages in their day to day life. In future, this work can be extended to other languages and code-mixed texts. Also, an ensemble model can be created by combining these pre-trained transfer learning models and performance can be evaluated.

References

Adoma, A. F., Henry, N.-M. and Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition, IEEE, pp. 117–121.

AlSalman, H. (2020). An improved approach for sentiment analysis of arabic tweets in twitter social media, IEEE, pp. 1–4.

Anand, T., Ramesh, K. and Singh, S. (2019). Out of the closet: Lexicon based sentiment analysis on tweets about homosexuality, IEEE, pp. 733–738.

Bhutani, B., Rastogi, N., Sehgal, P. and Purwar, A. (2019). Fake news detection using sentiment analysis, IEEE, pp. 1–5.

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. **URL:** *https://arxiv.org/abs/2109.00227*

Cheng, L. C. and Tsai, S. L. (2019). Deep learning for automated sentiment analysis of social media, Association for Computing Machinery, Inc, pp. 1001–1004.

Damas, C. A. and Mochetti, K. (2019). An analysis of homophobia on vandalism at wikipedia, IEEE, pp. 1–2.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Ghanghor, N., Ponnusamy, R., Kumaresan, P. K., Priyadharshini, R., Thavareesan, S. and Chakravarthi, B. R. (2021). IIITK@LT-EDI-EACL2021: Hope speech detection

for equality, diversity, and inclusion in Tamil , Malayalam and English, *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Association for Computational Linguistics, Kyiv, pp. 197–203.
URL: *https://aclanthology.org/2021.ltedi-1.30*

Goel, A. K. and Batra, K. (2020). A deep learning classification approach for short messages sentiment analysis, Institute of Electrical and Electronics Engineers Inc.

Ilmania, A., Abdurrahman, Cahyawijaya, S. and Purwarianti, A. (2018). Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis, IEEE, pp. 62–67.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mantoro, T., Merdianti, M. and Ayu, M. A. (2021). Sentiment analysis of the papuan movement on twitter using naïve bayes algorithm, Institute of Electrical and Electronics Engineers Inc.

Melton, J., Bagavathi, A. and Krishnan, S. (2020). Del-hate: A deep learning tunable ensemble for hate speech detection, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1015–1022.

Mittal, A. and Patidar, S. (2019). Sentiment analysis on twitter data: A survey, Association for Computing Machinery, pp. 91–95.

Rajapaksha, P., Farahbakhsh, R. and Crespi, N. (2021). Bert, xlnet or roberta: The best transfer learning model to detect clickbaits, *IEEE Access* **9**: 154704–154716.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* .

Vimali, J. S. and Murugan, S. (2021). A text based sentiment analysis model using bi-directional lstm networks, Institute of Electrical and Electronics Engineers Inc., pp. 1652–1658.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp. 6383–6389.
URL: *https://www.aclweb.org/anthology/D19-1670*

Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P. and Saini, S. (2020). Bi-lstm and ensemble based bilingual sentiment analysis for a code-mixed hindi-english social media text, Institute of Electrical and Electronics Engineers Inc.