Natural Language Processing for Equality, Diversity and Inclusion

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics
by Research

by

Ishan Sanjeev Upadhyay
2018114009
ishan.sanjeev@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June, 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Natural Language Processing for Equality, Diversity and Inclusion" by Ishan Sanjeev Upadhyay, has been carried out under my supervision and is not submitted elsewhere for a degree.

June 23, 2023	Adviser: Prof. Radhika Mamidi

To my family and friends

Acknowledgments

I would like to express my gratitude to my advisor, Prof. Radhika Mamidi for her constant support and guidance. She was always there to discuss my ideas and concerns and always made time for me. She also gave me freedom to explore my various research interests.

I would also like to thank the people who have collaborated with me on various research papers: Anshul Wadhawan, KV Aditya Srivatsa, Mukund Choudhary, Tathagata Raha, Vaibhav Bajaj and Priyanka Srivastava. I learned a lot from these collaborators and my journey would not have been successful without them.

I would like to thank all the professors who have taught me, I learned something from every course and each course left an impact on me and in some way, guided my research work.

Last but not the least, I would like to thank my parents. They have always been there for me and have helped me navigate through the various challenges I faced at IIIT-H.

Abstract

Social media has experienced significant growth in the past decade and has enabled people to connect with people all over the world, have increased access to information and have an opportunity to express themselves and join like-minded communities. However, hate speech and online harassment are significant problems, with around two-thirds of adults under 30 having experienced some form of online harassment. Therefore, it becomes essential to regulate content on social media and it needs to be done automatically due to the large volume of daily content.

In this thesis, we attempt to solve the above-mentioned problems by building best-in-class classifiers for novel datasets. One way to tackle harmful content online is to have a positive reinforcement approach and encourage positive and supportive messages. We propose a Hope Speech Detection model trained on a first-of-a-kind hope speech dataset. In the first approach, we used contextual embeddings to train classifiers using logistic regression, random forest, SVM, and LSTM based models. The second approach used a majority voting ensemble of 11 models obtained by fine-tuning pre-trained transformer models. Our model ranks first in terms of F1 score in the English language.

While supporting and boosting positive content online is helpful, there should also be a distinction made between content that is positive and content that seems positive but encourages emotion suppression. Over the past few years, there has been a growing concern around toxic positivity on social media, a phenomenon where positivity is used to minimize one's emotional experience. In this thesis, we create a dataset for toxic positivity classification from Twitter and an inspirational quote website. We then perform benchmarking experiments using various text classification models and show the suitability of these models for the task.

While there are many hate speech classifiers trained on a generic hate speech definition, there is a lack of datasets that focus on homophobia and transphobia. In this thesis, we describe our approach to classify homophobia and transphobia in social media comments. We used an ensemble of transformer based models to build our classifier. Our classifier ranks 1st in terms of F1 score.

Contents

Chapter Page

List of Figures

List of Tables

Chapter 1

Introduction

Social media has experienced significant growth in the past decade, with platforms such as Twitter, Facebook and Instagram becoming household names. According to Stastica [1], the number of global social media users is projected to reach 5.85 billion by 2027, up from 3.9 billion in 2020 (Figure 1.1). Social media has brought with it the ability to stay connected with people around the world. However, there are some problems that social media platforms face. This thesis will focus on the issue of harmful text on social media. More specifically we will develop classifiers and a dataset that promote inclusivity within online communities. We seek to accomplish this by identifying instances of hate speech, encouraging hopeful speech, and simultaneously addressing the issue of toxic positivity.

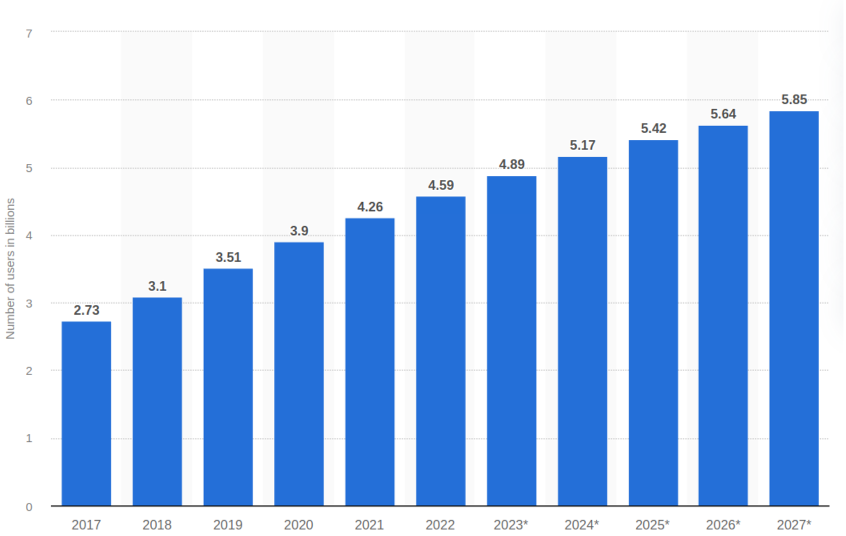


Figure 1.1: Growth of Social Media Users

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

1.0.1 Motivation

Let us delve into the benefits of social media and how the magnitude of these benefits is reduced by certain issues that social media platforms face today.

- **Connectivity and Communication**: The ability to connect with friends and family and people from all over the world is one of the key benefits of social media. This has made it easier to stay in touch with loved ones as well as make new friends and expand one's social circle [1] .

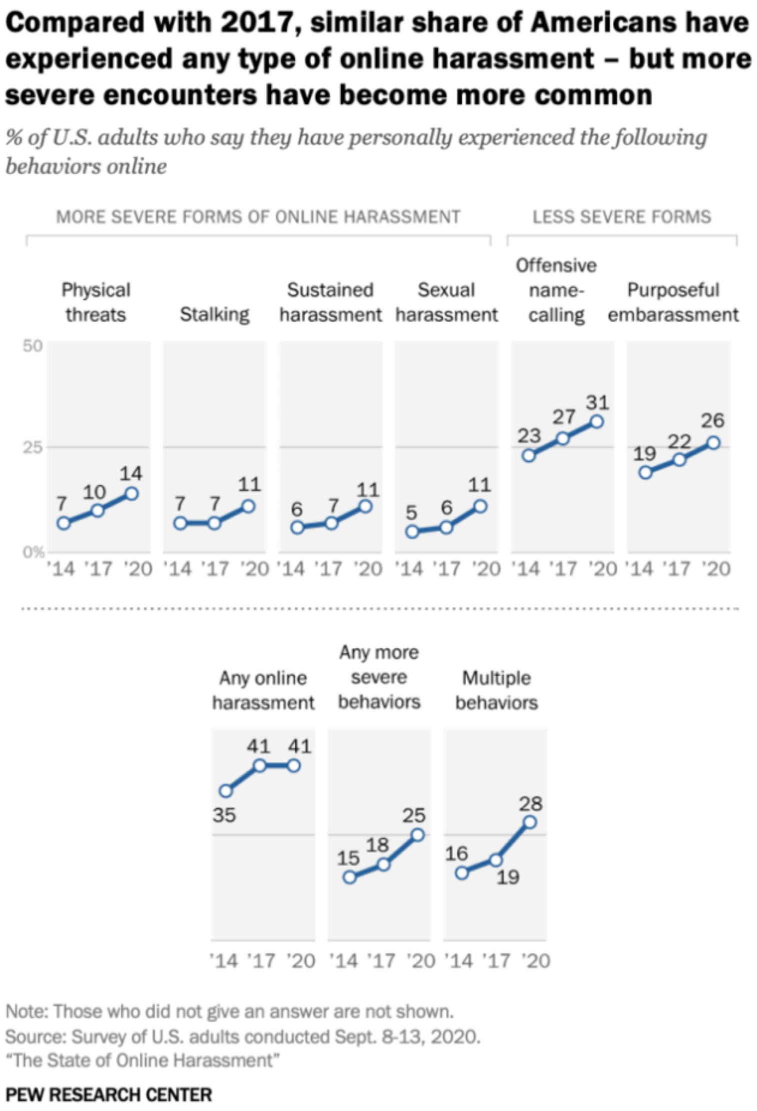
- **Greater Opportunity for Self Expression**: Individuals are provided with greater opportunities for self-expression. Partial or complete anonymity also allows users to express themselves more openly. Social media can help LGBT individuals by providing a platform for them to connect with others who share similar experiences and identities. This can help to assuage feelings of loneliness and isolation that these individuals might face in their offline communities and help them achieve community-building gratifications [2] [3]. Social media is also used to share information about events and community resources.

- **Increased Access to Information** : Social media has made it easier for people to access information on a wide range of topics. Social media platforms like LinkedIn and Twitter are often used to stay updated with news and events.

However, the benefits of social media come with its set of problems. Let us explore few of them.

- **Cyber-bullying and Hate Speech**: While social media allows for greater opportunity for self expression, it is also used negatively by users to spread hate and bully others, often with little or no repercussions [4]. It is seen that for LGBT individuals emotional investment in social media was negatively related to psychological well-being since users with strong emotional investment may place greater importance on how they are perceived online and may find it difficult to separate their online presence from their offline lives [2].

  According to a survey done by the Pew Research Center [5], 41% of Americans have personally experienced some form of online harassment. They also found that a growing share of Americans have reported experiencing more severe forms of harassment such as stalking, physical threats, sexual harassment and sustained harassment (Figure 1.2). **Social media is the most common venue for harassment**, with 75% of participants reporting that their most recent experience of online abuse was on social media.



Compared with 2017, similar share of Americans have experienced any type of online harassment – but more severe encounters have become more common

% of U.S. adults who say they have personally experienced the following behaviors online

Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

PEW RESEARCH CENTER

Figure 1.2: Growth of Severe Encounters of Online Harassment

Roughly two-thirds of adults under 30 have experienced any form of online harassment activities, making online harassment a particularly common feature of online life for young adults, with young adults being more susceptible to facing more serious harassing behaviours.

The survey also found that lesbian, gay or bisexual adults (LGB) are more likely than heterosexual adults to have experienced online harassment (Figure 1.3). Overall, 68% of LGB adults have experienced online harassment, compared with 39% of heterosexual adults. LGB adults are also more likely to have experienced more severe forms of online harassment, with 51% reporting that they have been targeted with severe behaviours, compared with 23% of heterosexual adults. While a small share of overall participants said that their harassment was due to their sexual orientation, 50% of lesbian, gay or bisexual adults who had been harassed online said that they thought it occurred because of their sexual orientation. Hate speech affects not just a person but has consequences for the entire group or society. In this thesis, we will be focusing on trying to solve the problem of hate speech to create more inclusive online communities.

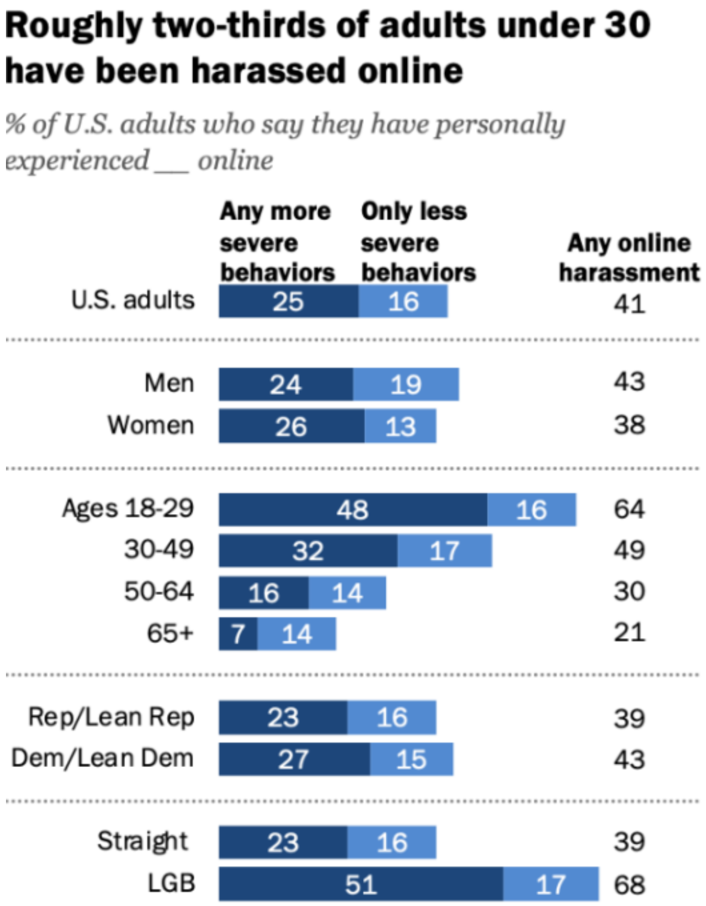


Figure 1.3: Percent of US Adults Who Say They Have Personally Experienced __ Online

- **Spread of misinformation**: The ease with which information can be spread allows for the spread of misinformation and fake news. With the volume of posts made every second, it becomes hard to administer every single post. Fake news detection systems have been made to deal with this.

A study done by Pew Research Center [6] showed that most Americans (64%) say that social media has a mostly negative effect on how things are going in the United States. Younger adults are more likely to say that social media has a positive impact but even in young adults between the ages of 18-29, the majority (54%) say that social media has a mostly negative effect. Hence due to the problems discussed above, it becomes vital to regulate, monitor and moderate social media, and it also becomes essential to do it automatically due to the large volume of content posted every minute on such platforms.

The public also favours stronger monitoring of posts that can be harmful. A Pew Survey [5] also showed that the public is highly critical of how social media companies tackle online harassment, with 79% of Americans saying that social media companies are doing an only a fair or poor job at addressing online harassment or bullying on their platforms. However, a minority of Americans back the idea of holding these platforms legally responsible for harassment on their sites. Only 33% of Americans say that people who have experienced harassment or bullying on social media sites should be able to sue the platforms on which it occurred.

The majority of U.S teens aged 13 to 17 prioritize a welcoming and safe online environment over people's ability to speak their minds freely online. It is no surprise that the generation that grew up with these social media platforms and that faces more severe online harassment also favours a more welcoming and safe environment online. Adults' views on the same issue are more divided, with half of adults prioritizing a welcoming and safe online environment and the other half valuing people's ability to speak their minds freely online. There are also disparities in views among different races and genders, with black adults and women more likely to prioritize a welcoming and safe online environment. [7].

Hence we can see the overall trend that ethnic and sexual minorities, young adults, and teens are more likely to face the effects and be more sensitive to online harassment and hate speech. At the same time, we see the support for better monitoring and regulation of such harmful posts on social media, with a majority of U.S teens prioritizing a welcoming and safe online environment over people's ability to speak their minds freely online. In this thesis, we will focus on trying to solve this problem of harmful text online and aim to create text classification solutions that foster more inclusive online communities. To that end, our work on homophobia and transphobia detection seeks to create a model that can help identify homophobic and transphobic speech so that vulnerable individuals can feel safe in online communities.

While content takedown is one of the ways of handling harmful speech, it has a censoring nature that conflicts with the democratic value of freedom of speech. Another method of dealing with this problem is 'counter speech' where users of the online community respond to harmful speech to stop it, ameliorate its impact or discourage it and support the person or group being attacked [8]. There has also been work on generating counter speech using text generation deep learning models [9]. Another way to deal with harmful speech is to take a positive reinforcement approach and encourage positive messages. In our work on hope speech detection, we create a model that can detect such positive speech so that it can be encouraged in online communities. While we want to promote positive and hopeful content, we do not want to encourage 'toxic positive' content, which results in one minimizing one's own negative feelings and suppressing negativity instead of acknowledging, processing and working through it. In this thesis, we introduce a dataset for toxic positivity detection and perform text classification using various transformer based models to establish the baseline results for this task.

1.1 Previous work

The most researched area in the field of classification of harmful text is ‘hate speech detection’. Hate speech is an umbrella term. It encompasses a wide range of harmful, offensive or discriminatory expressions. Data collection and classification are the two main parts of this kind of research. In this section we will see the various datasets and text classification models used.

1.1.1 Datasets

Reference	Data Source	Type of Classification	Number of Hate Labels	Size
10	Youtube, Facebook	Multi-label hateful language and targets	29	5,143
11	Youtube, Reddit	Binary and multi-lable hate speech	8	Binary: 988, Multi-lable:433
12	Twitter	Multi-label harassment	5	24,189
13	Twitter	Online harassment	1	30,000
14	Twitter	Toxic behaviour between high school students	1	16,901
15	Twitter	Multi-label hateful,abusive behaviour	2	80,000
16	Twitter	Muti-label East Asian Hate	3	20,000
17	Twitter	Hate against immigrants and women	3	19,600
18	Twitter	Ambivalent sexism	2	22,142

Table 1.1: Summary of datasets described in this section

Various datasets exist for hate speech detection; some focus on ‘general hate speech’ while others focus on a certain kind of hate speech. Datasets also originate from different sources such as social media websites or other sources like newspaper articles, blog posts etc. [10] created a dataset consisting of 5,143 texts extracted from Youtube and Facebook. The authors developed an extensive taxonomy and annotated the dataset with 29 hate categories, such as accusation, promoting violence, humiliation, swearing, specific nations, specific persons etc. It also considers both hateful language and targets. [11] created a dataset from Youtube and Reddit comments. The dataset has two variants: binary and multi-label. The multi-label dataset was annotated for 8 categories: Violence, Directed/Undirected, Gender, Race, National Origin, Disability, Sexual Orientation, and Religion. A total of 1421 sentences were annotated, with 988 for binary classification and 433 for multi-label classification. [12] created a dataset with 24,189 tweets sourced from Twitter. Since only a very small percentage of Tweets are hate speech, the authors developed a lexicon from online resources containing offensive words to narrow down the search for tweets that might have the presence of harassment. The lexicon covered five categories: sexual, racial, appearance-related, intellectual, political and a generic category that contains profane words not exclusively attributed to the five specific types of harassment. They then utilized the first five categories of the lexicon as seed terms for collecting Tweets. They annotated 24,189 tweets for the presence of harassment. [13] Created a dataset containing 35,000 tweets from Twitter annotated for online harassment. Certain hashtags and phrases were used to filter tweets like #whitePower, the jews, feminist, religion of hate etc. [14] created a dataset that captures toxic behaviour between high school students. For this, the authors first identified 143 high school students' twitter profiles and then expanded it by looking at the friends and follower list of each seed profile and applying some heuristics. They then collected tweets from these profiles and used the lexicon from [12] to filter tweets that would have a higher chance of containing toxic behaviour. This resulted in 456 accounts from which 688 interactions that consist of 16,901 tweets could be extracted. [15] conducted an eight month study of abusive behavior on Twitter and used a crowdsourcing methodology to annotate a collection of 80,000 tweets with abuse-related labels and performed statistical analysis merge and eliminated some labels resulting in a final set of labels. [16] Collected tweets from Twitters's streaming API using 14 hashtags that relate to East Asia and COVID-19. A total of 159,320 unique tweets were collected this way. From this dataset, the 1000 most used hashtags were annotated by three annotators and assigned the stance towards the Asian Entity ranging from very negative to very positive. Through this, 97 hashtags marked as negative and very negative by at least one anotators were identified and 10,000 tweets were collected that used one of these 97 hashtags. Another 10,000 tweets were taken from the 159,320 tweets at random. Each tweet was also assigned to 5 mutually exclusive categories. The authors create a 20,000 tweets dataset with 5 labels with 3 focusing on hate. There is also a dataset created for SemEval [17] for the task of hate speech against immigrants and women detection. That consisted of 13,000 Tweets in English and 6,600 tweets in Spanish. [18] created a dataset for the detection of ambivalent sexism by collecting tweets from Twitter and using ambivalent sexism theory to annotate the tweets into three categories: benevolent, hostile and others.

As we can see from the datasets described above (Summary in Table 1.1), datasets can be sourced from various places, with Twitter being the most popular platform due to its API. Furthermore, we can see that while some datasets focus on binary classification, others focus on multi-label and more fine-grained classification. We can also see that while some datasets concentrate on a more general definition of hate speech, online harassment or toxic behaviour, others focus on more specific issues like hate against immigrants, women, East-Asians or ambivalent sexism.

1.1.2 Text Classification

Detecting hate speech or harmful text can be seen as a text classification problem in machine learning. More specifically, this is a sentiment analysis problem since we are trying to determine the emotional tone of a piece of text. While the earliest methods for sentiment analysis relied on rule-based approaches like using lexicons and seeing the presence of certain words to classify text, by the early 2000s, machine learning techniques like naive bayes and support vector machines became more popular

Soon, deep learning models started being used and showed remarkable improvements in various NLP tasks, including sentiment analysis. Deep learning models can learn the complex relationships between the sentiment labels and text data and extract features from the text data automatically. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, started being used for sentiment analysis and achieved state-of-the-art results.

Most recently, Transformer based models have gained popularity. The Transformer architecture was introduced in 2017 in the paper “Attention Is All You Need” by Vaswani et al. [19] The Transformer model uses self-attention mechanisms to capture long-range dependencies between words in a sentence, which allows it to process text more efficiently than traditional RNNs and LSTMs. The application of Transformer models to sentiment analysis has led to significant improvements in accuracy and efficiency. For example, the BERT (Bidirectional Encoder Representations from Transformers) model has achieved state-of-the-art results on several benchmark datasets for sentiment analysis. BERT is a pre-trained language model that uses a Transformer-based architecture to encode contextual information in text data.

The following is a brief description of the models we will be using in this thesis.

- **Support Vector Machine** [20] is a machine learning algorithm that can be used for sentiment analysis. The key idea behind SVMs is to find the hyperplane that best separates the positive and negative examples in the feature space, a high-dimensional space where each example is represented as a vector of features. Frequencies of n-grams, part of speech, and presence of certain words could all be possible features. The hyperplane is chosen such that it maximizes the margin between the positive and negative examples. The margin is the distance between the hyperplane and the closest positive and negative examples. By maximizing the margin, SVMs seek to find a decision boundary that is as generalizable as possible, meaning it can accurately classify new examples that were not seen during training.

- **Convolutional Neural Networks** [21] are a type of deep learning model that have been applied to a wide range of computer vision and natural language processing tasks, including sentiment analysis.

  In a CNN, the input data is typically represented as a matrix or tensor, where each element of the matrix corresponds to a feature of the input, such as a pixel value in an image or a word embedding in text. The CNN architecture consists of several layers of convolutions, pooling, and non-linear activations.

The convolutional layer is the core component of a CNN, and performs a feature extraction operation on the input data. The convolution operation is repeated across the entire input data, resulting in a set of feature maps that capture local patterns in the input data.

The output of the convolutional and pooling layers is typically fed into one or more fully connected layers, which perform a classification or regression task on the extracted features.

To represent the text data as input to the CNN, various techniques can be used, such as the use of pre-trained word embeddings, such as Word2Vec or GloVe, which represent each word as a vector in a high-dimensional space. The word embeddings are then fed into the input layer of the CNN, where they are convolved with filters to extract local patterns in the input data.

- **Long Short-Term Memory** [22] is a type of RNN that seeks to solve the short-term memory or vanishing gradient problem that RNNs face. In an LSTM, the hidden state is replaced by a cell state, which is updated based on the current input and the previous cell state. The cell state is controlled by three gates: the input gate, the forget gate, and the output gate.

  The input gate determines how much new information should be added to the cell state based on the current input. The forget gate determines how much information from the previous cell state should be discarded. The output gate determines how much of the cell state should be used to compute the output at the current time step.

  The gates are controlled by sigmoid activation functions, which output values between 0 and 1. These values are used to scale the input, forget, and output vectors, allowing the LSTM to selectively store and retrieve information over long periods of time.

- **BERT** [23] is a type of pre-trained language model that has been widely used in natural language processing tasks, including sentiment analysis.

  BERT is a neural network model that is trained on large amounts of text data to learn general language representations. The model is based on the transformer architecture.

  The pre-training process for BERT involves two stages: masked language modeling (MLM) and next sentence prediction (NSP). In MLM, a certain percentage of the input tokens are randomly masked, and the model is trained to predict the original value of the masked tokens. In NSP, the model is trained to predict whether two input sentences are consecutive or not.

  Once the pre-training process is complete, the BERT model can be fine-tuned for a specific task, such as sentiment analysis. To do this, the final layer of the BERT model is replaced with a task-specific layer, which is trained on a smaller dataset of labeled examples.

  To represent the input text as input to BERT, various techniques can be used, such as tokenization, which involves breaking the text into individual tokens or subwords. The tokens are then mapped to their corresponding embeddings, which are learned during the pre-training process.

- **RoBERTa**[24] is based on the same transformer architecture as BERT, but with a few key differences. For example, RoBERTa uses dynamic masking, which involves randomly masking out tokens at each training epoch, rather than the static masking used in BERT. This helps the model to learn more robust representations of the input data.

  In addition, RoBERTa uses a larger training corpus and a longer training schedule than BERT. The RoBERTa model is trained on a dataset of over 160 GB of text data, which is significantly larger than the 13 GB dataset used to train the original BERT model. The training schedule is also longer, with RoBERTa being trained for 100 epochs compared to 40 epochs for BERT.

- **ALBERT**[25] was designed to address some of the limitations of the BERT model, such as its computational cost and memory requirements.

  The main innovation of ALBERT is its use of a factorized embedding parameterization, which allows the model to share its parameters across different layers. This reduces the number of parameters in the model, making it more efficient and easier to train.

  In addition, ALBERT uses cross-layer parameter sharing, which allows the model to share parameters across different layers of the network. This further reduces the number of parameters in the model and improves its performance on downstream natural language processing tasks.

  ALBERT is pre-trained on a large corpus of text data, using the same masked language modeling and next sentence prediction tasks as BERT. Once pre-training is complete, the model can be fine-tuned for specific tasks, such as sentiment analysis, by replacing the final layer with a task-specific layer and training on a smaller dataset of labeled examples.

1.2 Thesis Contribution

In this thesis, we aim to create classifiers and a dataset that make online communities more inclusive by classifying and diminishing hate speech, incentivizing hope speech while at the same time keeping a check on toxic positivity. The major contributions of our work are:

- We proposed a Hope Speech Detection model that ranked first with an F1 score of 0.93 .

- We proposed a Homophobia and Transphobia detection model that ranked first in terms of weighted F1 score (0.94) and second in terms of macro-F1 score.

- We created a new dataset from Twitter and an inspirational quote website for Toxic Positivity Detection and described the annotation procedure, and established baseline results.

- We compared various transfomer-based models and experimented with ensembling techniques such as majority voting and weighted ensemble using XGBoost Random Forest Classifier.

1.3 Thesis Workflow

- In chapter 2, we describe the approach taken to solve the task of Hope Speech Detection.
- In chapter 3, we describe our new dataset to classify ‘Toxic Positivity’ and the baseline models established.
- In chapter 4, we describe the approach taken to solve the task of Homophobia and Transphobia Detection.
- Chapter 5 concludes the thesis with a discussion of future work.

Chapter 2

Hope Speech Detection

As we discussed in the previous chapter, The spread of hate speech on social media is a major problem. While there have been attempts made at hate speech detection [26, 27] to stop the spread of negativity, this form of censorship can also be misused to obstruct rights and freedom of speech. Furthermore, hate speech tends to spread faster than non-hate speech [28].While there has been a growing amount of marginalized people looking for support online [29, 30], there has been a substantial amount of hate towards them too [31]. Therefore, detecting and promoting content that reduces hostility and increases hope is important. Hope speech detection can be seen as a rare positive mining task because hope speech constitutes a low percentage of overall content [32]. There has been work done on hope speech or help speech detection before that has used logistic regression and active learning techniques [32, 33]. In this chapter, we will be doing the hope speech detection task on the HopeEDI dataset [34, 35] which consists of user comments from Youtube in English, Tamil and Malayalam.

We will first look at the task definition, followed by the methodology used. We will then look at the experiments and results followed by conclusion and future work.

2.1 Task Definition

The given problem is a comment level classification task for the identification of“hope speech” within YouTube comments, wherein they are to be classified as“Hope speech”,“Not hope speech” and “Not in intended language”. The data provided in the task was annotated at a per-comment basis wherein a comment could be composed of more than one sentence.

2.2 Methodology

This section talks about the methodology that we have used to solve the task. As shown in Figure 1, the pipeline involves preprocessing, language detection, transliteration (for Indian languages), and hope speech detection. These steps are described in this section.

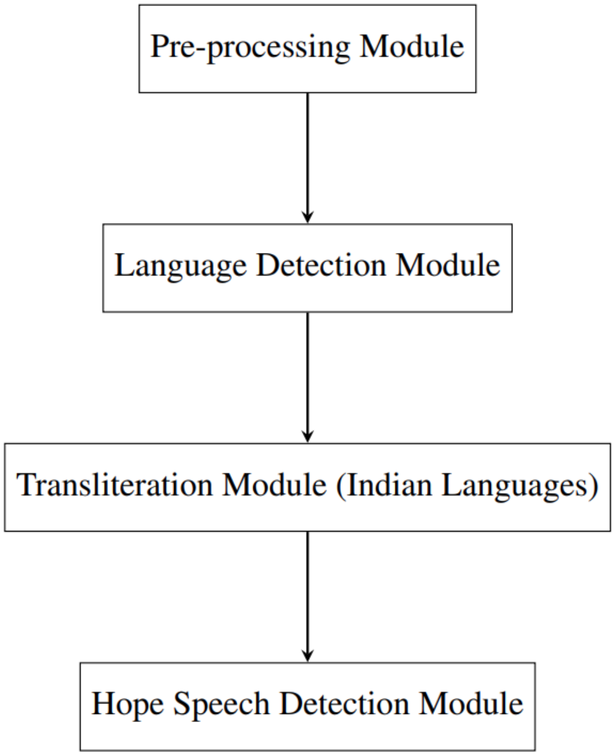

Pre-processing Module
Language Detection Module
Transliteration Module (Indian Languages)
Hope Speech Detection Module

Figure 2.1: Methodology Pipeline

2.2.1 Pre-processing Module

The preprocessing module involved the following:

- Removing special characters and excess whitespaces
- Removing emojis
- Make text lowercase.

These steps were taken to make the text more uniform. Special characters like "@" and "#" were removed because they did not serve as good features for classification and language detection. Emojis were removed because they were sparsely used in the dataset.

2.2.2 Language Detection Module

The task involves classifying text into hope, not-hope and not-language. Language detection module marks the not-language sentences. We use Google's language detection library [36] to do this. The

Tamil and Malayalam datasets are code-mixed. Inter-sentential, intra-sentential, tag code-mixing and code-mixing between Latin and native script is observed in the Tamil and Malayalam datasets. Google's language detection library does not work on such code-mixed data. Since the Tamil and Malayalam sentences involve code-mixed data, language detection can not be done on them using the Google language detection library. We observed that sentences that were marked as not-Tamil and not-Malayalam were mostly English sentences with some of them being Hindi and other languages. Hence, we adopted a heuristic where we marked sentences as not-Tamil or not-Malayalam if the sentences were detected to be in English or Hindi, other sentences were assumed to belong to the respective language.

2.2.3 Transliteration Module

After language detection, sentences that are classified to be in Tamil and Malayalam undergo transliteration. Tamil and Malayalam text have code-mixing between Latin and native script, hence transliteration is done to make the entire text in the native script. This step is also important because it makes the text closer to the kind of text IndicBert is trained on. Transliteration was done by using the indic-transliteration library [1].

2.2.4 Hope Speech Detection Module

After preprocessing and transliteration (for Indian languages), the text is sent to the hope speech detection module. The hope speech detection module is responsible for predicting if a text is hope speech or not hope speech. We have used the following for our experiment.

2.2.4.1 Models

A more detailed description can be found in the introduction section of this thesis.

BERT [23] is based on the transformer architecture. Using its multi-layer encode module, It is able to jointly utilize both left and right contexts across all layers to pre-train its bidirectional representations. We have fine tuned "bert-base-uncased" model on the dataset for one of our experiments.

RoBERTa [24] is a transformer architecture which is based on optimizations made to the BERT approach. It trains on more data and bigger batches, removes next sentence prediction objective that BERT used, trains on longer sequences and introduces dynamic masking (ie. mask tokens change during training epochs). We fine-tuned the "roberta-base" model on the provided data. The roberta-base model is trained on 160 GB of English text from five different datasets.

ALBERT [25] is a transformer architecture based on BERT but with fewer parameters. We used IndicBERT [37] which is a multilingual ALBERT model pre-trained on 12 major Indian languages. We also fine-tuned "albert-base-v2" model for our experiment in English.

[1] https://github.com/sanskrit-coders/indic_transliteration

LSTM Long Short-Term Memory [22] networks seek to solve the short-term memory or vanishing gradient problem that RNNs face. They do so by having internal gates that regulate the flow of information. Information flows through a mechanism known as cell states. The cell can make decisions about what to store, what to forget and what the next hidden state should be.

Random Forest Classifier Random forests [38] use an ensemble of a large number of decision trees generally trained with the bagging method. These decision trees are created using random subsamples of the given dataset with replacement (bootstrap dataset) and a random subset of the features. New samples are classified by choosing the prediction made by most decision trees (majority voting).

Support Vector Machine Support vector machine (SVM) [20] is a supervised learning method that can be used for classification or regression. We have used SVM for classification. The objective of the SVM classification algorithm is to find the hyper-plane that most accurately differentiates two classes that have been plotted on a f dimensional plane where f is the number of features.

Logistic Regression Logistic regression [39]is a statistical model used for binary classification. It does so by using a logistic function to model the binary outcome. It can be extended for multiclass classification problems.

2.2.4.2 Ensemble Process

Ensembles can help make better predictions by reducing the spread of predictions. Hence, lowering variance and improving accuracy. We used a voting based ensemble method where we trained N models on N different training and validation data obtained by random shuffling. We then chose the majority voting as the merging technique to produce our final prediction y. In majority voting, the final prediction y is decided based on which prediction is made by the majority of the models . We made two ensembles, one each of 7 models and 11 models and chose the ensemble that gave the best weighted F1 score.

2.3 Experiments

Initially, the entire database is preprocessed to remove extra tab spaces, punctuations, emojis, mentions and links. In the case of Malayalam and Tamil, we also transliterate the entire database. Then we distributed our experimentation procedure into two different approaches. In the first approach, we finetune our pre-trained masked language models using the train and validation splits for the purpose of making them more suitable to the subsequent classification task. Thereafter, contextual embeddings for each sentence in the dataset are produced by calculating the average of the second to last hidden layer for every single token in the sentence. We then trained Logistic Regression, Random Forest, SVM and RNN-based classifier models using these embeddings. In the second approach, all the sentences are encoded into tokens using the respective tokenizers and then we add a linear layer on top of the pre-trained model layers after dropout. All the layers of the devised model are then trained such that the error is back propagated through the entire architecture and the pre-trained weights of the model are modified to reflect the new database. For both these approaches, we then calculated predictions for the test split and reported performance metrics. For English, we try out three different pre-trained models: “roberta-base”,“bert-base-uncased”, and “albert-base-v2” for both the approaches. For Tamil and Malayalam however, only the IndicBERT model is applicable for either approach.

Language	Database	Hope	Not Hope	Other Lang.
English	Train	1962	20778	22
	Dev	242	2569	2
Tamil	Train	6327	7872	1961
	Dev	757	998	263
Malayalam	Train	1668	6205	691
	Dev	190	784	96

Table 2.1: Data distribution by class

2.3.1 Dataset

The HopeEDI dataset consists of Youtube comments marked as “hope”, “not hope” and “other language” in three languages: English, Tamil and Malayalam. The distribution of hope, not hope and other language tag in the training and development datasets is shown in table 2.1. The ratio of hope to not hope is around 0.09 in English, 0.26 in Malayalam and 0.79 in Tamil. Table 2.2shows the data distribution between training, development and test datasets. There are a total of 28,451 comments in English, 10,705 comments in Malayalam and 20,198 comments in Tamil. Data in Tamil and Telugu has code-mixing. In the English dataset, there are instances where English comments are annotated as not English. For example, “Fox News is pure Garbage!” is annotated as not English in the training set. This contributes some noise to the English dataset.

	English	**Tamil**	**Malayalam**
Training	22762	16160	2564
Development	2843	2018	1070
Test	2846	2020	1071
Total	28451	20198	10705

Table 2.2: Data distribution by language

2.3.2 System Settings

In the first approach, we run the task of masked language modelling on our database for 4 epochs for each of the 5 model-database combinations. Afterwards, the sentence input token length is limited to 512 and the embeddings extracted by evaluation on the input sequences by the model are of length 768. The RNN based classifier is composed of an LSTM layer and two dense layers. In the second approach, the encoded sentences are in the form of a data loader class, containing the respective input IDs and attention masks, with a batch size of 16. These are then passed into a model that implements a dropout of 30% and the output from the final linear layer is used for classification.

2.3.3 Evaluation Metrics

We used F1 and weighted F1 scores for evaluating our model.

$$F1\ Score = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

Weighted F1 scores are calculated by taking the F1 scores for each label and then doing a weighted average by the number of true instances of each label.

$$weighted\ F1 = \frac{(F1_i \times y_i + F1_j \times y_j)}{(y_i + y_j)}$$

y_i and y_j are the number of true instances of class i and class j respectively and $F1_i$ and $F1_j$ are the F1 scores of class i and j respectively.

2.3.4 Results

Our experimentation involved two approaches. In the first approach, we used contextual embeddings (E) to train classifiers using logistic regression (LR), random forest (RF), SVM, and LSTM based models. In the second approach we used an ensemble of 11 models which were generated by fine-tuning (FT) pre-trained transformer models after adding an output layer. We used majority voting to get our

Model	Method Used	Macro Precision	Weighted Precision	Macro Recall	Weighted Recall	Macro F1-Score	Weighted F1-Score
BERT	E + LR	0.778	0.911	0.656	0.924	0.695	0.913
	E + RF	0.834	0.902	0.526	0.916	0.528	0.881
	E + SVM	0.771	0.86	0.489	0.866	0.488	0.837
	E + LSTM	0.456	0.833	0.500	0.913	0.477	0.871
	FT	0.759	0.915	0.728	0.915	0.742	0.915
ALBERT	E + LR	0.703	0.881	0.538	0.912	0.549	0.883
	E + RF	0.832	0.900	0.506	0.914	0.491	0.874
	E + SVM	0.456	0.833	0.500	0.913	0.477	0.871
	E + LSTM	0.657	0.878	0.571	0.905	0.591	0.887
	FT	0.755	0.916	0.705	0.924	0.725	0.919
RoBERTa	E + LR	0.794	0.914	0.657	0.926	0.700	0.915
	E + RF	0.840	0.905	0.535	0.917	0.544	0.885
	E + SVM	0.821	0.899	0.517	0.915	0.512	0.878
	E + LSTM	0.791	0.918	0.693	0.928	0.729	0.921
	FT	0.753	0.915	0.748	0.922	0.745	**0.923**

Table 2.3: Metrics for English language [E: Contextualized Embeddings, LR: Logistic Regression, FT: Finetuned model, RF: Random Forest]

final prediction. Results for both the approaches on our test split generated from the provided train and dev datasets are reported in Tables 2.3, 2.4 and 2.5 for English, Tamil and Malayalam respectively. We report the macro-averaged and weighted recall, precision and F1-score for each possible model-method combination. While the weighted F1 scores are more representative of how well a model performs, the disparity between the weighted and macro-averaged scores demonstrates how disproportionate a certain model's effectiveness is in predicting the different classes. For English, the second approach involving finetuning is the best performing one for each of the models tested, closely followed by the Logistic Regression and LSTM-based methods in the first approach. The roberta-base model seems to have a slight edge over the other two tested models. For Tamil and Malayalam, the second approach is still the best performer, but by a greater margin.

Model	**Method Used**	**Macro Precision**	**Weighted Precision**	**Macro Recall**	**Weighted Recall**	**Macro F1-Score**	**Weighted F1-Score**
Indic -BERT	E + LR	0.473	0.482	0.484	0.520	0.441	0.464
	E + RF	0.511	0.516	0.506	0.544	0.458	0.482
	E + SVM	0.278	0.309	0.500	0.556	0.357	0.397
	E + LSTM	0.591	0.587	0.501	0.557	0.364	0.403
	FT	0.635	0.637	0.627	0.636	0.623	**0.629**

Table 2.4: Metrics for Tamil language [E: Contextualized Embeddings, LR: Logistic Regression, FT: Finetuned model, RF: Random Forest]

Model	**Method Used**	**Macro Precision**	**Weighted Precision**	**Macro Recall**	**Weighted Recall**	**Macro F1-Score**	**Weighted F1-Score**
Indic -BERT	E + LR	0.645	0.729	0.501	0.790	0.447	0.699
	E + RF	0.395	0.623	0.499	0.788	0.440	0.696
	E + SVM	0.386	0.610	0.492	0.777	0.433	0.683
	E + LSTM	0.367	0.579	0.471	0.745	0.413	0.652
	FT	0.776	0.842	0.743	0.842	0.756	**0.837**

Table 2.5: Metrics for Malayalam language [E: Contextualized Embeddings, LR: Logistic Regression, FT: Finetuned model, RF: Random Forest]

2.4 Conclusion and Future Work

We presented our approach for hope speech detection in English, Tamil, and Malayalam on the HopeEDI dataset. We used two approaches. The first approach involved using contextual embeddings to train various classifiers. The second approach involved using a majority voting ensemble of 11 models which were obtained by fine-tuning pre-trained transformer models. The second approach using the roberta-base model was the best performing model for English, giving a weighted F1 score of 0.93. The second approach using IndicBERT model gave the best performance for Tamil and Malayalam, giving a weighted F1 score of 0.75 for Malayalam and 0.49 for Tamil. In the future, we plan to fine-tune transformers pre-trained on code mixed data. Data augmentation methods like synonym replacement and random insertion could be used to fine-tune the model on more data.

Chapter 3

Toxic Positivity Detection

3.1 Introduction

In the previous chapter we discussed how encouraging positive and hopeful content in online communities can be one of the ways of dealing with harmful speech. However, one needs to be careful while defining what is positive speech because speech that may look positive may actually be ‘toxic positive’. Toxic positivity can be defined as the overgeneralization of a positive state of mind that encourages using positivity to suppress and displace any acknowledgement of stress and negativity [40, 41]. The popularity of the term “toxic positivity” peaked during the COVID 19 pandemic (refer to figure 3.1) where it was used to identify advice that focused on just looking at the positive at a time when people were hurting due to loss of life, loss of jobs and other traumatic events.

Toxic positivity results in one minimizing one’s own negative feelings and suppressing negativity instead of acknowledging, processing and working through it. Some examples of toxic positivity include telling someone to focus on the positive aspects of a loss, telling someone that positive thinking will solve all their problems, suggesting that things could be worse and shaming someone for expressing negative emotions. This suppression of emotions is not only unhelpful but also leads to poorer recovery from the negative effects of the emotion. Accepting and working through one’s emotions is the better route to take while dealing with negative emotions [42].

Macro level events like COVID 19 and climate change disasters have distressed many people in the past few years [43]. Social media is used by people having mental health issues or going through a tough time to find community, support, advice and encouraging messages [44]. However, it becomes important to be able to differentiate between messages that may help uplift an individual and those that may look positive but promote suppression of emotions and cause great harm in the long term recovery from negative emotions. The harms of toxic positivity are not only limited to its deleterious mental health outcomes but it can also be used to uphold oppression by making people ignore the oppression that is going on and encouraging them to “just be positive”.

We aim to create a dataset for toxic positivity and perform text classification using various transformer based models to establish the baseline results for this task.

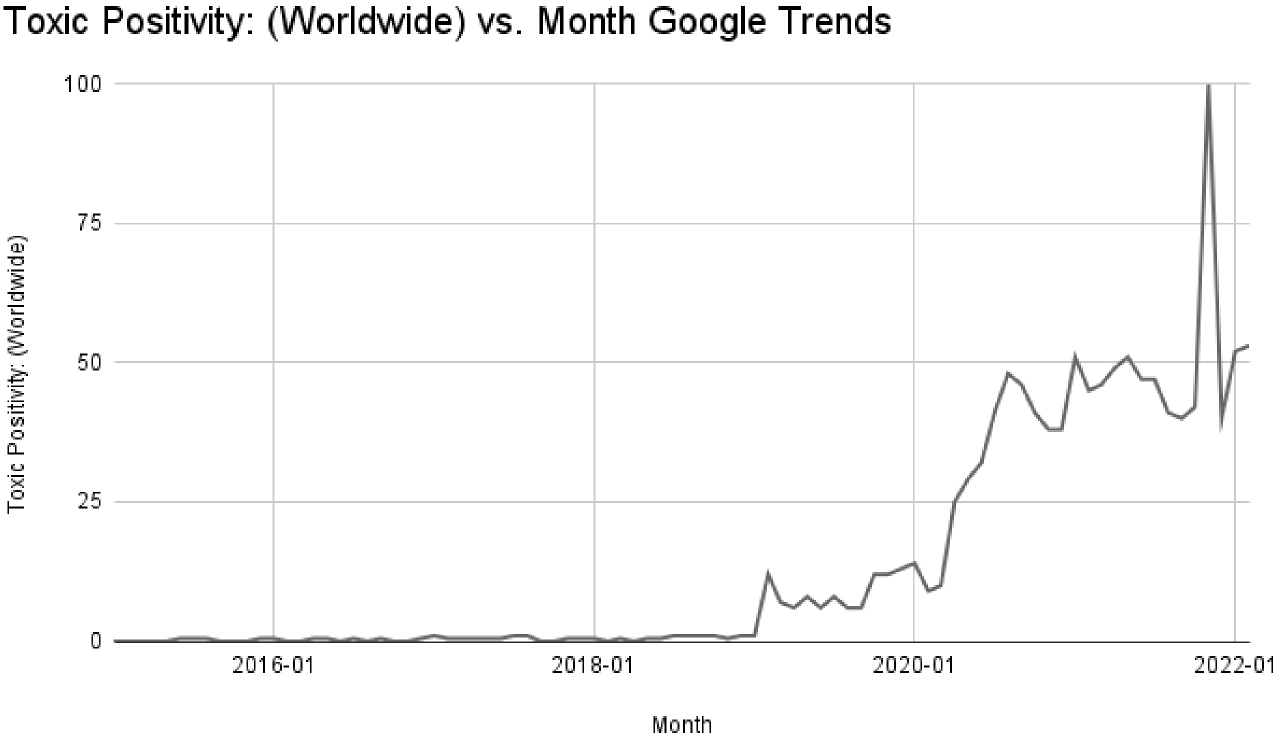


Figure 3.1: Worldwide Google Trends showing search interest of the term ”Toxic Positivity”.

3.2 Related Work

There have been studies that show the ineffectiveness and deleterious effects of emotion suppression. Gross and John (2003) [45] showed that people who suppressed their emotions had a greater experience of negative emotions while also expressing lesser positive emotion. They also showed that using suppression is related negatively to well being. A study done by Campbell-Sills et al. (2006)[42] involved dividing 60 participants diagnosed with anxiety and mood disorders into two groups. One group was given a rationale for suppressing their emotions while the other was given a rationale for accepting emotions. It was found that suppression was ineffective in reducing distress while watching an emotion-provoking film. It was also seen that the suppression group showed a poorer recovery from the changes in negative affect after watching the film compared to the acceptance group. A similar observation is seen in the case of physical pain as well. Cioffi and Holloway (1993) [46] divided participants into three groups during a cold-pressor pain induction (CPT) where participants would dip their hands in cold water for as long as tolerable. The first group was told to pay attention to the pain, the second was told to focus on their room at home as a distraction, and the third was told to suppress the sensations they felt. It was seen that the group that focused on the pain had a faster recovery from the pain and the suppression group had the slowest recovery from pain. Suppressing pain has shown to have negative outcomes, while accepting it is observed to be as a better strategy. Ford et al. (2018)[47] through longitudinal and lab studies showed that habitually accepting mental experiences broadly predicted psychological health and that it reduced negative emotional response and experience. Hence toxic positivity, with its overemphasis on thinking positively and having a positive state of mind, encourages emotion suppression rather than emotional acceptance which has negative consequences for the person who engages in it.

Lecompte-Van Poucke (2022) [48] conducted a critical discourse analysis of toxic positivity as a discursive construct on Facebook. Two corpora of posts from organizations that promoted endometriosis awareness (an invisible chronic condition) were analyzed using systematic functional linguistics, pragma-dialectics and critical theory. The study showed that users on social media platforms often engage in toxic positivity or forced positive discourse which is inspired by the neoliberal "positive thinking" ideology, leading to a less inclusive online community.

As we discussed in chapter 1, in the field of NLP, there have been many papers focusing on hate speech detection using support vector machine (SVM), long short term memory networks (LSTM),convolutional neural network (CNN), transformers and other machine learning models [49, 50, 51, 52]. These works use Twitter posts (tweets) to create datasets. YouTube and Reddit comments have also been used in some works [11, 53]. There have been recent efforts in hope speech detection as well [54]. The HopeEDI dataset [34] is a hope speech dataset that contains Youtube comments that have been marked for hope and not-hope speech. We have discussed in the previous chapter, the classification work that we did on this dataset that achieved the best result in the English language. [55].

However, to the best of our knowledge, there has been no prior work on creating datasets and classification models for toxic positivity.

3.3 Dataset Creation

3.3.1 Data Extraction and Pre-processing

We sourced our data from two sources. Twitter and inspirational quote website BrainyQuote[1] which is one of the largest quotation websites.

The reason for sourcing data from BrainyQuotes was that we observed that a lot of motivational quotes being shared on Twitter were ones that were said by famous personalities. Hence, including popular quotes from a quotation website is helpful. We made a web scraper using Beautiful Soup 4[2] library in python to extract a subset of quotations from the website.

For the Twitter data, we extracted tweets using Twitter API [3] we queried using hashtags like #MondayMotivation to #SundayMotivation and hashtags like #InspirationalQuotes, #Motivation, #SelfLove

[1] http://www.brainyquote.com

[2] BeautifulSoup Documentation

[3] Twitter API Documentation

and #AdviceForSuccess. We also took quotes from widely followed inspirational or motivational twitter accounts.

After collecting the data, pre-processing was performed. Bylines of quotes were removed because it was not useful information for annotation and to also to ensure that there was no annotator bias. For tweets, hashtags and ”@” tags were removed. The Twitter data and BrainyQuotes data was also manually filtered to remove sentences that were not inspirational, motivational or advisory in nature. Examples of the kind of data removed are given in Table 3.2. A total of 4,250 quotes and tweets were collected for annotation after the data elimination and pre-processing steps.[4].

3.3.2 Dataset Annotation

Two annotators annotated the data for toxic positivity. The annotators were linguistics students. An annotation workshop was conducted for the annotators where they were sensitized to the topic of toxic positivity through academic works as described in the related works section and examples of toxic positivity. The annotators were then asked to annotate 50 sentences separately and then their annotator agreement was measured and was found to have a Kappa score of 0.72.We used Cohen’s Kappa coefficient to calculate Inter Annotator Agreement [56] . The annotators then discussed their disagreements and came to a better understanding of the annotation guidelines. They annotated another 50 sentences and got a better Kappa score of 0.76. They again had a discussion about their disagreements. After this exercise, they were told to annotate the dataset separately without communicating with each other. The 100 sentences used for training the annotators were discarded and are not a part of this dataset of 4,250 sentences. It was observed that sentences that had the following general characteristics were marked as toxic positive:

- Encouraging hiding or suppressing negative emotions.
  - Example: “A negative mind will never give you a positive life.”
- Encouraging focusing on positivity rather than processing negative emotions.
  - Example: “Every time I hear something negative, I will replace it with a positive thought.”
- Minimizing someone’s negative feelings.
  - Example:“You cannot be lonely if you like the person you’re alone with.”

A few categories of sentences or quotes we emerged when were studying the dataset. We decided to annotate for them as well. The categories of the sentences were as follows.

- **Worldview**: sentences that are philosophical, abstract and provide an insight into the worldview of the writer. Example: “Things may come to those who wait, but only the things left by those who hustle”

[4] Dataset Link

- **Personal Experience**: sentences that provide insights based on the writer's personal experience. Example: "I always did something I was a little not ready to do. I think that's how you grow. When there's that moment of 'Wow, I'm not really sure I can do this,' and you push through those moments, that's when you have a breakthrough."

- **Advice**: sentences that are more instructional in nature and provide straightforward recommendations and advice. Example: "Do one thing every day that scares you."

- **Affirmation**: First-person sentences that are used as affirmations. Example: "I choose to make the rest of my life, the best of my life."

The same annotators annotated the categories of sentences as well. The same process of annotating 100 sentences, 50 sentences at a time and discussing disagreements was followed to train the annotators.

3.3.3 Dataset Statistics

Out of the 4,250 sentences, 512 were annotated as toxic positive, which constitutes 12% of the dataset.The rest of the 3738 sentences were non-toxic positive. Examples of toxic and non-toxic positive sentences are presented in Table 3.1.

Worldview was the most common category of sentence occurring 73.6% of the time with advice occurring 16.7% of the time and the rest occurring less than 10% of the time in the dataset. Exact figures are presented in Table 3.4.

It was also seen that 44% of the sentences that belonged to the affirmation category were toxic positive. 21% of the sentences belonging to the advice category were toxic positive, while 14% and 8% of sentences belonging to the personal experience and the worldview category respectively were toxic positive. We noticed that in our dataset, most affirmation sentences were focused on emotion suppression, and hence they were marked as toxic positive. The non-toxic positive affirmations focused on gratitude, having a growth mindset and self-acceptance, although they were fewer in number.

We got a Kappa score of 0.82 for the toxic positivity (toxic or non-toxic) annotation and a Kappa score of 0.74 for category annotations (worldview, advice, personal experience or affirmation).

3.4 Methodology

We used the following transfomer-based models for text classification, a more detailed description of these models is provided in chapter 1.

Sentence	**Class**
When people say there is a ’reason’ for the depression, they insult the person who suffers, making it seem that those in agony are somehow at fault for not ’cheering up.’ The fact is that those who suffer - and those who love them - are no more at fault for depression than a cancer patient is for a tumor.	Non-Toxic Positive
Just like it’s not healthy to think overly negative thoughts, exaggeratedly positive thoughts can be equally detrimental. If you overestimate how much of a positive impact a particular change will have on your life, you may end up feeling disappointed when reality doesn’t live up to your fantasy.	Non-Toxic Positive
Do what you feel in your heart to be right	Non-Toxic Positive
The secret of getting ahead is getting started.	Non-Toxic Positive
Being positive is like going up a mountain. Being negative is like sliding down a hill. A lot of times, people want to take the easy way out, because it’s basically what they’ve understood throughout their lives.	Toxic Positive
You must not under any pretense allow your mind to dwell on any thought that is not positive, constructive, optimistic, kind.	Toxic Positive
While you’re going through this process of trying to find the satisfaction in your work, pretend you feel satisfied. Tell yourself you had a good day. Walk through the corridors with a smile rather than a scowl. Your positive energy will radiate. If you act like you’re having fun, you’ll find you are having fun.	Toxic Positive
You can’t live a positive life with a negative mind and if you have a positive outcome you have a positive income and just to have more positivity and just to kind of laugh it off.	Toxic Positive

Table 3.1: Examples of toxic positive and non-toxic positive sentences in the dataset.

Removed Text	**Source**
Check out this new print for SPRING! #SpringForArt #ThisSpringBuyArt #gardeners #gardens #Inspire #InspirationalQuotes	Twitter
A future Metaverse, a social network for the people by the people, around jobs and finance in the decentralised world.Tomorrow's job fair in 3 dimensions at your fingertips. #MondayMotivation #cryptocurrency #blockchain #Crypto #jobseeker #Trader #Jobs #trading #ICO	Twitter
The failure of Lehman Brothers demonstrated that liquidity provision by the Federal Reserve would not be sufficient to stop the crisis; substantial fiscal resources were necessary.	BrainyQuote
Museums are managers of consciousness. They give us an interpretation of history, of how to view the world and locate ourselves in it. They are, if you want to put it in positive terms, great educational institutions. If you want to put it in negative terms, they are propaganda machines.	BrainyQuote

Table 3.2: Examples of the text removed during dataset creation.

Class	**Number of sentences**
Toxic Positive	512
Non-Toxic Positive	3738

Table 3.3: Distribution of toxic positive and non-toxic positive sentences.

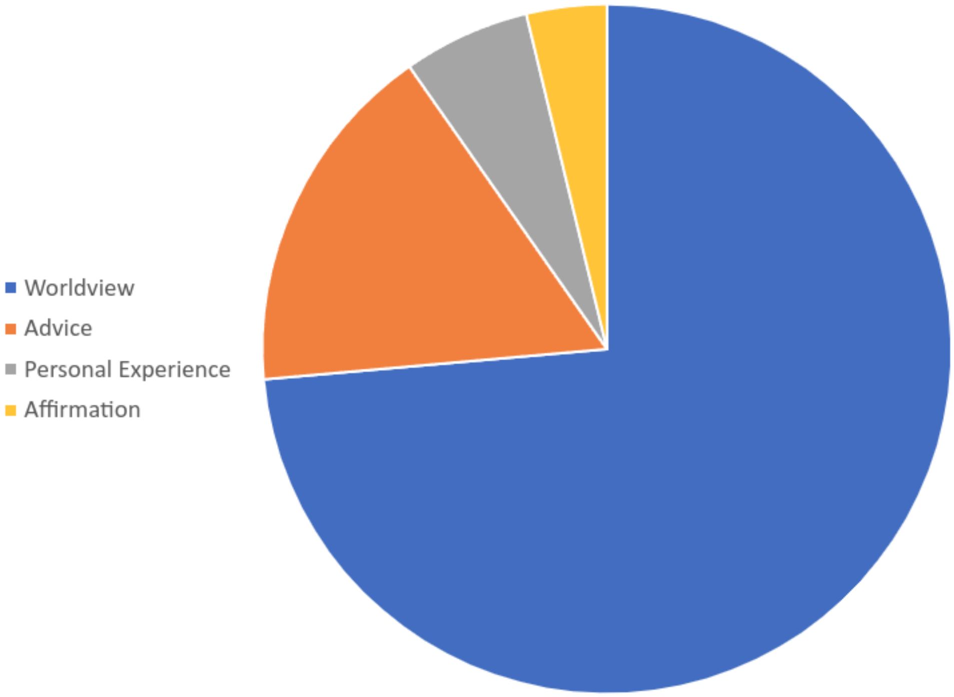


Figure 3.2: Sentence Category Distribution Pie Chart.

- **BERT**: BERT [23] is a transformer encoder with several encoder layers, each with several self-attention heads. We have fine-tuned the “bert-base-uncased” model in our implementation.

- **RoBERTa**: RoBERTa [24] is a transformer-based encoder built by modifying the original BERT architecture. It utilizes more data with longer average sequence lengths and larger batches. For our classifier, we have fine-tuned the “roberta-base” model.

- **ALBERT**: ALBERT [25] is yet another transformer encoder based on BERT but aimed at being lighter than its predecessor. We have fine-tuned the “albert-base-v2” model in our implementation.

We also experimented with an ensemble based classifier for which we additionally used the following:

- **XGBoost Random Forest Classifier**: Random Forest Classifiers [57] are widely used for ensemble classification. They consist of a large number of decision trees, each set to only a subset of the overall feature-set of the data. This helps create numerous weak learners with relatively low correlation. The majority verdict of these weak learners tends to outperform an individual predictor tasked with the entire feature-set. We have made use of the implementation of the Random Forest Classifier by XGBoost [58].

- **Bayesian Optimization**: Bayesian Optimization [59] is a sequential global optimization strategy for various black-box functions and is used for models across Machine Learning. It attempts

Type of sentence	**Number of sentences**
Worldview	3128
Advice	709
Personal Experience	253
Affirmation	160

Table 3.4: Distribution of the various types of sentences occurring in the dataset.

to determine the prior distribution of the system (i.e model hyperparameters), which yields the optimal posterior distribution (i.e objective function) by iteratively testing the prior and updating the posterior accordingly. It provides a more computationally efficient yet fine-grained search space than more exhaustive methods such as grid search. In our work, Bayesian optimization is used for tuning the hyperparameters (i.e. number of tree estimators, train subsample ratio, and column subsample ratio) of the Random Forest Classifier. We make use of the implementation by the bayesian-optimization Python library [60].

3.5 Experiments and Results

We experimented with 3 transformer models BERT, RoBERTa, and ALBERT. Each of the classification models utilizes a pretrained Transformer encoder, i.e. BERT-Base, RoBERTa-Base, and ALBERT-Base. The pooled output layer from each encoder is passed through respective dropout layers ($p = 0.3$) for further regularization and linear layers (mapping from a vector size of 768 to the number of classification categories, i.e. 2). A softmax function is applied to each of the size-2 vectors for normalized likelihoods of the two classes. The results from these models are provided in Table 3.5.

We also experimented with an ensemble-based classifier. The classifier is an ensemble of three predictors with a random forest classifier on top (as shown in Figure 3.3). The predictors were the three text classification transformer based models as mentioned above.

The likelihoods from each of the predictors were concatenated and passed as features to an XGBoost Random Forest Classifier to generate an ensemble class prediction. After a Bayesian Search for the classifier parameters on the validation set, the number of tree estimators w set to 149, subsample ratio of the training samples to 0.50, and subsample ratio of columns for each split to 0.33.

Each of the Transformer encoder predictors were trained using AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), with Cross Entropy loss, using a linear training scheduler. The encoder pipelines were trained with an initial learning rate of $2e^{-5}$ and the XGBoost ensemble classifier with a learning rate of 1.0. The predictors were trained for 6 epochs . The predictions from the epoch with the best

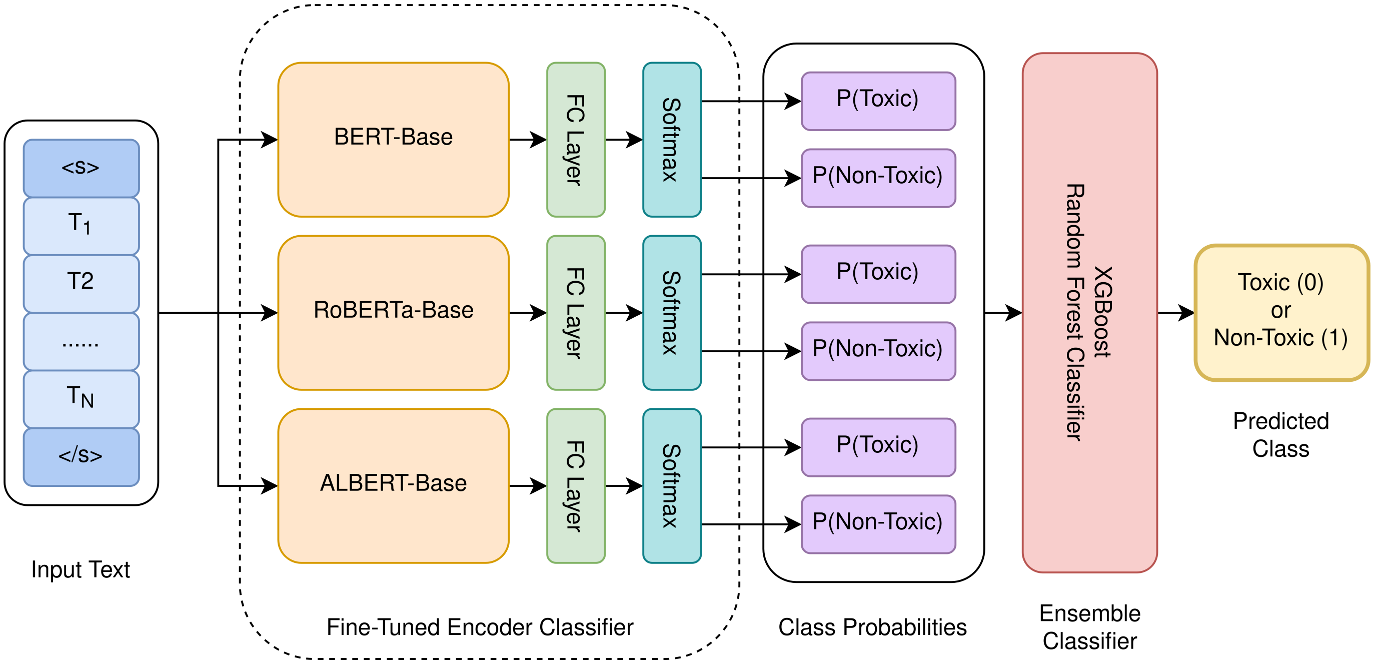


Figure 3.3: Schematic overview of the model architecture.

validation weighted macro F1 score were utilized for the ensemble classification. The overall batch size for the pipeline was set to 16.

The ensemble model generalized better than the individual models producing the highest macro F1 score of 0.71 and a weighted F1 score of 0.85 as seen in Table 3.5. As the toxic tweets comprise of only a small portion of the data (14.5%), models performing well on non-toxic tweets tend to have inflated weighted-F1 scores. Therefore we opted for macro-F1 as the main performance metric for this task.

3.6 Conclusion and Future Work

In this work, we created a dataset for toxic positivity detection. We scraped 4,250 sentences from Twitter and the inspirational quote website BrainyQuote. We then annotated them and achieved a Kappa score of 0.82 for toxic positivity classification. We then performed experiments using transformer-based models for text classification. Our ensemble model gave us the best results achieving a macro F1 score of 0.71 and a weighted F1 score of 0.85. As more people turn to social media to get help when they are going through a tough time, it becomes important for them to be able to differentiate between positive and toxic positive messages. Furthermore, being able to recognize toxic positivity is also important for chatbots and other automated systems that aim to provide mental health assistance. We hope that our work contributes to further research in this field. In the future, we plan to extend the study by introducing

Model	**Macro Precision**	**Weighted Precision**	**Macro Recall**	**Weighted Recall**	**Macro F1**	**Weighted F1**
BERT	0.78	0.84	0.6	0.86	0.63	0.83
RoBERTa	0.71	0.85	0.7	0.84	0.68	0.85
ALBERT	0.71	0.83	0.65	0.85	0.67	0.84
Ensemble	**0.76**	**0.85**	**0.69**	**0.86**	**0.71**	**0.85**

Table 3.5: Classification results of various models used on the dataset.

a larger dataset in English as well as other languages. We also plan to look at toxic positivity from a discourse perspective and annotate for toxic positivity on online discussion forums.

Chapter 4

Ensembled Transformers Against Homophobia and Transphobia

4.1 Introduction

Social media platforms allow people from all walks of life to connect with each other. However, as we discussed in the first chapter,abusive and hateful content on these platforms can take a psychological toll on its users [61] [62]. Lesbian, gay, bisexual and transgender individuals are more vulnerable to mental illness as compared to their heterosexual peers [63] [64] [65]. Content takedowns still remain the most popular method of maintaining inclusive online communities. Hence, it becomes even more important to be able to detect such hateful content for vulnerable individuals.

There has been a lot of work done in the domain of hate speech detection [66] [67]. There has also been work on hate speech intervention [68]. Shared tasks like SemEval 2019 Task 6 have focused on identifying and categorizing offensive language on social media [69]. Datasets for this task have been created in multiple languages as well. bohra-etal-2018-dataset created a Hindi-English code mixed text dataset for hate speech detection from tweets on Twitter. Mubarak et al. (2021) [70] created a 1000 tweets Arabic dataset for offensive language detection with special tags for vulgarity and hate speech. Sigurbergsson and Derczynski (2020)[71] created a Danish hate speech detection dataset containing 3600 user generated comments social media websites. There have been datasets created for Greek [72] and Turkish [73] as well. Chakravarthi et al.(2021)[74] created a code-mixed Tamil,Malayalam and Kannada dataset for offensive language identification. Support vector machines, long short-term memory networks, convolutional neural networks and now transformer based architectures have been used to detect hate speech. However, there has not been much work in trying to specifically identify homophobic or transphobic text.

In this paper, we will describe our approach for classifying transphobic and homophobic comments in the dataset provided by Chakravarthi et al. (2021)[75] as a part of the shared task on homophobia and transphobia detection in social media comments [76].

Language	**Number of comments**	**Number of tokens**	**Number of characters**
English	4,946	82,111	438,980
Tamil	4,161	197,237	539,559
Tamil-English	6,034	66,731	435,890
Total	15,141	346,079	1,414,429

Table 4.1: Distribution of comments in English, Tamil and Tamil-English.

Class	**English**	**Tamil**	**Tamil English**
Homophobic	276	723	465
Transphobic	13	233	184
Non-anti-LGBT+ content	4,657	3,205	5,385
Total	4,946	4,161	6,034

Table 4.2: Distribution between Homophobic, Transphobic and Non-anti-LGBT+ content.

4.2 Dataset Description

The dataset consists of a total of 15,141 comments in 3 languages: English, Tamil and Tamil-English code-mixed (refer to Table 4.1 for data distribution). Each comment has one of three labels ”Homophobic”, ”Transphobic” and ”Non-anti-LGBT+ content” (label distribution in Table 4.2).

4.3 Methodology

In this section we will describe the models used in our experimentation.A more detailed description is given in the introduction section of the thesis.

- **BERT**: BERT [23] is a Transformer-based language model. It consists of layered encoder units, each with a self-attention layer followed by fully-connected layers.For this task, we have used the pretrained bert-base-uncased model from HuggingFace [77].
- **RoBERTa**: RoBERTa [24] is a Transformer-based language model which improves upon the BERT architecture along several metrics offered by the GLUE benchmark [78]. It is not trained

on the NSP task and involves dynamic masking for the MLM task. It is also trained over a much larger dataset with longer sentence lengths. For this task, we have used the pretrained roberta-base model.

- **HateBERT** [79] is a re-trained BERT model to detect abusive language in English. It is trained on large amounts of banned Reddit comments extracted from the RAL-E dataset. It has been shown to outperform the BERT model in several hate-speech detection tasks.

- **IndicBERT**: IndicBERT [80] is an ALBERT Transformer encoder [25] finetuned on data from 12 major Indian languages, including 549M tokens of Tamil. Despite having significantly lower parameters than other multilingual encoders such as mBERT [23] or XLM-R [81], it outperforms them on several metrics of the IndicGLUE benchmark [80]. We have used the IndicBERT model as a TLM for the Tamil and Tamil-English tracks.

- **XGBoost Random Forest Classifier**: Random Forest Classifiers [57] are meta estimators which consist of numerous decision trees, each fit upon a subset of features from a subset of rows of the data. The ensemble of many such weak learners tends to outperform a single large decision tree. The low correlation between the constituent trees also provides for more feature coverage and curbs over-fitting. For this task, we use XGBoost's implementation of Random Forest Classifiers [58].

- **Bayesian Optimization**: The aim of any hyperparameter optimization strategy is to find the hyperparameter set which fetches the best value over the object function. Bayesian Optimization [59] is an iterative optimization algorithm that aims to minimize the number of hyperparameter sets that must be evaluated before arriving at the optimal distribution. It has been shown to generate optimal solutions in significantly fewer iterations than traditional methods such as grid search. For this task, we have used the Python library: bayesian-optimization [60].

4.4 Experiments and Results

The only pre-procesesing step done on the dataset before training was the change of emojis to text using the demoji library in python [1]. Our pipeline comprises an ensemble of several Transformer-based language models (TLM), namely: BERT, RoBERTa, and HateBERT for the English track and IndicBERT for the Tamil and Tamil-English tracks. Three copies of each TLM are used with different parameter initializations in each track. This allows for the copies to capture different features of the data. In addition to this, for each track, a layer of attention is applied to each constituent encoder layer outputs of the TLMs. This is necessary since each layer captures a different kind of information, which are variably relevant for our task. The weighted and combined output from the attention layer is then

[1] https://pypi.org/project/demoji/

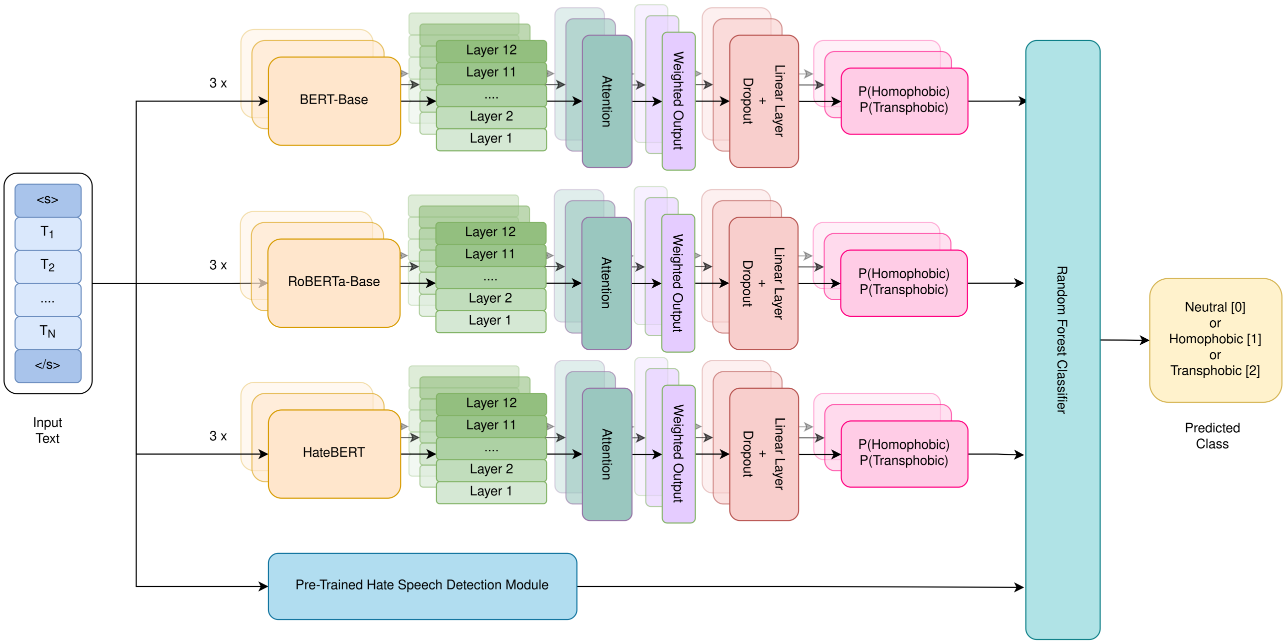


Figure 4.1: Schematic overview of the architecture of our model.

passed through a final linear layer and dropout layer ($p = 0.3$), followed by a Softmax operation to generate the predicted probabilities of detecting homophobic content in the given input text.

In the English track, we also use a pretrained hate-speech detection model implemented on HuggingFace [77]. Architecturally, is a ByT5-Base model [82] finetuned on HuggingFace's tweets_hate_speech_detection dataset [83]. Figure 4.1 provides a schematic overview of the architecture of our model.

The prediction probabilities are generated by each model of a track are passed as input features to a Random Forest Classifier. This helps further optimize our predictions by weighing the importance of the different architectures for the task.

Each of the TLM pipelines was finetuned upon Cross Entropy loss using AdamW optimizer [84] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with an initial learning rate of $2e^{-5}$ for 6 epochs each using a linear scheduler. The epoch checkpoint with the highest validation F1 score was selected for further use. The hyperparameters of the Random Forest Classifier were estimated using 10 seeds and 100 iterations of Bayesian Optimization. The ensemble classifier was trained with a learning rate of 1.0.

As can be seen in Table 4.3, our ensemble model performed better than the individually trained models giving a macro F1 score of 0.49 which was the 2nd highest macro F1 score in the shared task. This model also had the highest weighted F1 score in the task. The IndicBERT ensembles trained on the Tamil and Tamil-English dataset give us a macro F1 score of 0.55 and 0.35 and a weighted F1 score of 0.86 and 0.83 respectively (refer Table 4.4).

Model	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**
BERT	0.92	0.48	0.42	0.44	0.9	0.92	0.91
RoBERTa	0.93	0.64	0.36	0.36	0.93	0.94	0.9
HateBERT	0.94	0.56	0.43	0.47	0.92	0.94	0.92
Ensemble	**0.94**	**0.52**	**0.47**	**0.49**	**0.93**	**0.94**	**0.94**

Table 4.3: Classification results of various models used on the English dataset.

Model	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**
Tamil-English	0.83	0.34	0.35	0.35	0.82	0.83	0.83
Tamil	0.88	0.52	0.58	0.55	0.85	0.88	0.86

Table 4.4: Classification results of IndicBERT finetuned on the Tamil-English and Tamil dataset.

4.5 Conclusion and Future Work

In this chapter, we described our approach for homophobia and transphobia detection in English, Tamil and Tamil-English. We used an ensemble of three transformer based models along with a pre-trained hate detection model to do the classification for English. Our model was ranked 2nd for the English classification task. For the Tamil and Tamil-English dataset three copies of the IndicBERT model was used to make our ensemble based model. The models placed 8th and 10th for Tamil and Tamil-English model respectively.

In the future, we can use data augmentation methods like paraphrasing and back translation to increase the diversity and quantity of homophobic and transphobic text. We can also incorporate transliteration into the pipeline for Tamil-English code mixed text since IndicBERT is not trained on code mixed text. We could also try to finetune transformers pre-trained on code mixed data.

Chapter 5

Conclusion

This thesis presents classification systems for three different kinds of sentiment analysis tasks. In the second chapter, we presented our approach for hope speech detection in English, Tamil and Malayalam. We tried two different methods. In the first method, we used contextual embeddings to train several different types of classifiers, including logistic regression, random forest, support vector machines, and LSTM-based models. In the second method, we used an ensemble of 11 models created by fine-tuning pre-trained transformer models and adding an output layer. The second approach using the roberta-base model was the best performing model for English, giving a weighted F1 score of 0.93.

While it is important to encourage positive messages online, it becomes vital to be able to differentiate between messages that may help uplift an individual and those that may look positive but promote suppression of emotions and cause great harm in the long-term recovery from negative emotions. To help with this problem, we created a dataset for toxic positivity detection. We scraped 4,250 sentences from Twitter and the inspirational quote website BrainyQuote. We then annotated them for toxic positivity. A few categories of sentences emerged when studying the dataset, and we annotated that too. The categories are worldview, personal experience, advice and affirmation. We observed that around 12% of the dataset was toxic positive. We then performed experiments using transformer-based models for text classification. Our ensemble model gave us the best results achieving a macro F1 score of 0.71 and a weighted F1 score of 0.85.

We also created a classifier for homophobia and transphobia. We use a collection of Transformer-based language models (TLM) for our pipeline, which includes BERT, RoBERTa, and HateBERT for the English track, as well as IndicBERT for the Tamil and Tamil-English tracks. Each track uses three copies of each TLM, each with different parameter initializations, to capture different features of the data. We apply a layer of attention to each constituent encoder layer output of the TLMs for each track since each layer captures different information relevant to our task. The weighted and combined output from the attention layer is then passed through a final linear layer and dropout layer (with a dropout probability of 0.3), followed by a Softmax operation to generate the predicted probabilities of detecting homophobic content in the input text.

For the English track, we also incorporate a pretrained hate-speech detection model from HuggingFace. The predicted probabilities generated by each model in a track are used as input features for a Random Forest Classifier to optimize our predictions by weighing the importance of the different architectures for the task. Our model ranked first in F1 score and second in macro F1 score for this task.

There remains scope for more work in this domain. Datasets can be made available in more languages which will help in more classifiers being built for hope speech and homophobia transphobia detection. Multi-modality can also be incorporated by looking at tweets with pictures in them. The toxic positivity dataset can be further extended with conversational data being taken into account. The discourse between social media users in terms of toxic positivity can be further studied. Chatbots that aim to provide mental health assistance can be made sensitive to toxic positivity.

Related Publications

1. **Ishan Sanjeev Upadhyay**, KV Aditya Srivatsa, and Radhika Mamidi. "Towards Toxic Positivity Detection." *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, July 2022

2. **Ishan Sanjeev Upadhyay**, KV Aditya Srivatsa, and Radhika Mamidi. "Sammaan@LT-EDI-ACL2022: Ensembled Transformers Against Homophobia and Transphobia.", *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, April 2021

3. **Ishan Sanjeev Upadhyay**, Nikhil E, Anshul Wadhawan, Radhika Mamidi. "Hopeful Men@LT-EDI-EACL2021: Hope Speech Detection Using Indic Transliteration and Transformers.", *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, May 2022

Other Publications

1. Tathagata Raha, **Ishan Sanjeev Upadhyay**, Radhika Mamidi, Vasudeva Varma. "IIITH at SemEval-2021 Task 7: Leveraging transformer-based humourous and offensive text detection architectures using lexical and hurtlex features and task adaptive pretraining" *Proceedings of the 15th International Workshop on Semantic Evaluation*, August 2021

2. Vaibhav Bajaj, Kartikey Pant, **Ishan Sanjeev Upadhyay**, Srinath Nair and Radhika Mamidi. "TEASER: Towards Efficient Aspect-based SEntiment Analysis and Recognition", *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, September 2021

Bibliography

[1] Andrew Perrin. Social media usage: 2005-2015, Oct 2015.

[2] Randolph C.H. Chan. Benefits and risks of lgbt social media use for sexual and gender minority individuals: An investigation of psychosocial mechanisms of lgbt social media use and well-being. *Computers in Human Behavior*, page 107531, Oct 2022.

[3] Stephan A. Brandt and Cheryl L. Carmichael. Does online support matter? the relationship between online identity-related support, mattering, and well-being in sexual minority men. *Computers in Human Behavior*, 111:106429, Oct 2020.

[4] Robert Slonje and Peter K. Smith. Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, 49(2):147–154, Apr 2008.

[5] Emily A. Vogels. The state of online harassment. *Pew Research Center*, Jan 2021.

[6] Brooke Auxier. 64% of americans say social media have a mostly negative effect on the way things are going in the u.s. today, Oct 2020.

[7] Colleen McClain. More so than adults, u.s. teens value people feeling safe online over being able to speak freely, Aug 2022.

[8] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1), Jan 2022.

[9] Mana Ashida and Mamoru Komachi. *Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions*. 2022.

[10] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), Jun 2018.

[11] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex Intelligent Systems*, Jan 2022.

[12] Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. A quality type-aware annotated corpus and lexicon for harassment research. *Proceedings of the 10th ACM Conference on Web Science*, May 2018.

[13] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, and Piyush Ramachandran. A large labeled corpus for online harassment research. *Proceedings of the 2017 ACM on Web Science Conference*, Jun 2017.

[14] Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. Alone: A dataset for toxic behavior among adolescents on twitter. *Lecture Notes in Computer Science*, page 427–439, 2020.

[15] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), Jun 2018.

[16] Bertie Vidgen, Austin Botelho, David A. Broniatowski, E. Guest, M. Hall, H. Margetts, Rebekah Tromble, Zeerak Waseem, and Scott A. Hale. Detecting east asian prejudice on social media, 2020.

[17] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*. 2019.

[18] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, Aug 2017.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[20] Marti A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998.

[21] Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, Cambridge, MA, USA, 1998.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[25] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[26] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics.

[27] Younghun Lee, Seunghyun Yoon, and Kyomin Jung. Comparative studies of detecting abusive language on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[28] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA, 2019. Association for Computing Machinery.

[29] Kris Gowen, Matthew Deschaine, Darcy Gruttadara, and Dana Markey. Young adults with mental health conditions and social networking websites: Seeking tools to build community. *Psychiatric Rehabilitation Journal*, 35(3):245–250, 2012.

[30] Zijian Wang and David Jurgens. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[31] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 85–94, New York, NY, USA, 2017. Association for Computing Machinery.

[32] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Hope speech detection: A computational analysis of the voice of peace, 2020.

[33] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, Jaime G. Carbonell, Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):454–462, Apr. 2020.

[34] Bharathi Raja Chakravarthi. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[35] Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, 2021.

[36] Nakatani Shuyo. Language detection library for java, 2010.

[37] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November 2020. Association for Computational Linguistics.

[38] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[39] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[40] Laura Sokal, Lesley Eblie Trudel, and Jeff Babb. It's okay to be okay too. why calling out teachers' "toxic positivity" may backfire. *EdCan*, 60, 2020.

[41] Eva Bosveld. *Positive Vibes Only: The Downsides of a Toxic Cure-All*. 2021.

[42] Laura Campbell-Sills, David H. Barlow, Timothy A. Brown, and Stefan G. Hofmann. Effects of suppression and acceptance on emotional responses of individuals with anxiety and mood disorders. *Behaviour Research and Therapy*, 44(9):1251–1263, Sep 2006.

[43] Donatella Marazziti, Paolo Cianconi, Federico Mucci, Lara Foresi, Chiara Chiarantini, and Alessandra Della Vecchia. Climate change, environment pollution, covid-19 pandemic and mental health. *Science of The Total Environment*, page 145182, Jan 2021.

[44] Kris Gowen, Matthew Deschaine, Darcy Gruttadara, and Dana Markey. Young adults with mental health conditions and social networking websites: Seeking tools to build community. *Psychiatric Rehabilitation Journal*, 35(3):245–250, 2012.

[45] James J. Gross and Oliver P. John. Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2):348–362, 2003.

[46] Delia Cioffi and James Holloway. Delayed costs of suppressed pain. *Journal of Personality and Social Psychology*, 64(2):274–282, 1993.

[47] Brett Q. Ford, Phoebe Lam, Oliver P. John, and Iris B. Mauss. The psychological health benefits of accepting negative emotions and thoughts: Laboratory, diary, and longitudinal evidence. *Journal of Personality and Social Psychology*, 115(6):1075–1092, Dec 2018.

[48] Margo Lecompte-Van Poucke. "you got this!": A critical discourse analysis of toxic positivity as a discursive construct on facebook. *Applied Corpus Linguistics*, 2(1):100015, Apr 2022.

[49] Bin Wang, Yunxia Ding, Shengyan Liu, and Xiaobing Zhou. Ynu_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 191–198. CEUR-WS.org, 2019.

[50] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, pages 745–760, Cham, 2018. Springer International Publishing.

[51] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, November 2019. Association for Computational Linguistics.

[52] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[53] Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA, 2020. Association for Computing Machinery.

[54] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Hope speech detection: A computational analysis of the voice of peace. *ECAI 2020*, page 1881–1889, 2020.

[55] Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on*

Language Technology for Equality, Diversity and Inclusion, pages 61–72, Kyiv, April 2021. Association for Computational Linguistics.

[56] Joseph L. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, Oct 1973.

[57] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[58] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[59] Jonas Mockus. *Bayesian approach to global optimization*. Kluwer Academic, 1989.

[60] Nogueira Fernando. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014.

[61] Michał Wypych and Michał Bilewicz. Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among ukrainian immigrants in poland. *Cultural Diversity and Ethnic Minority Psychology*, Jan 2022.

[62] Brendesha M. Tynes, Michael T. Giang, David R. Williams, and Geneene N. Thompson. Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health*, 43(6):565–569, Dec 2008.

[63] S E Gilman, S D Cochran, V M Mays, M Hughes, D Ostrow, and R C Kessler. Risk of psychiatric disorders among individuals reporting same-sex sexual partners in the national comorbidity survey. *American Journal of Public Health*, 91(6):933–939, Jun 2001.

[64] Michael P. Marshal, Laura J. Dietz, Mark S. Friedman, Ron Stall, Helen A. Smith, James McGinley, Brian C. Thoma, Pamela J. Murray, Anthony R. D'Augelli, and David A. Brent. Suicidality and depression disparities between sexual minority and heterosexual youth: A meta-analytic review. *Journal of Adolescent Health*, 49(2):115–123, Aug 2011.

[65] Sari L Reisner, Ralph Vetters, M Leclerc, Shayne Zaslow, Sarah Wolfrum, Daniel Shumer, and Matthew J Mimiaga. Mental health of transgender youth in care at an adolescent urban community health center: a matched retrospective cohort study. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 56(3):274–9, 2015.

[66] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd.

[67] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1), Mar 2016.

[68] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China, November 2019. Association for Computational Linguistics.

[69] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[70] Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

[71] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France, May 2020. European Language Resources Association.

[72] Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France, May 2020. European Language Resources Association.

[73] Çağrı Çöltekin. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France, May 2020. European Language Resources Association.

[74] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv, April 2021. Association for Computational Linguistics.

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

[76] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, May 2022.

[77] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.

[78] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

[79] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, August 2021. Association for Computational Linguistics.

[80] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November 2020. Association for Computational Linguistics.

[81] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[82] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *CoRR*, abs/2105.13626, 2021.

[83] Roshan Sharma. tweets-hate-speech-detection - datasets at hugging face, 2019.

[84] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.