# HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion

**Bharathi Raja Chakravarthi**
Insight SFI Research Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
bharathiraja.akr@gmail.com

## Abstract

Over the past few years, systems have been developed to control online content and eliminate abusive, offensive or hate speech content. However, people in power sometimes misuse this form of censorship to obstruct the democratic right of freedom of speech. Therefore, it is imperative that research should take a positive reinforcement approach towards online content that is encouraging, positive and supportive contents. Until now, most studies have focused on solving this problem of negativity in the English language, though the problem is much more than just harmful content. Furthermore, it is multilingual as well. Thus, we have constructed a Hope Speech dataset for Equality, Diversity and Inclusion (HopeEDI) containing user-generated comments from the social media platform YouTube with 28,451, 20,198 and 10,705 comments in English, Tamil and Malayalam, respectively, manually labelled as containing hope speech or not. To our knowledge, this is the first research of its kind to annotate hope speech for equality, diversity and inclusion in a multilingual setting. We determined that the inter-annotator agreement of our dataset using Krippendorff's alpha. Further, we created several baselines to benchmark the resulting dataset and the results have been expressed using precision, recall and F1-score. The dataset is publicly available for the research community. We hope that this resource will spur further research on encouraging inclusive and responsive speech that reinforces positiveness.

## 1 Introduction

With the expansion of the Internet, there has been substantial growth all over the world in the number of marginalised people looking for support online (Gowen et al., 2012; Yates et al., 2017; Wang and Jurgens, 2018). Recently, due to the lockdowns enforced as a consequence of the COVID-19 pandemic, people have started to look at online forums as an emotional outlet when they go through a tough time. The importance of the online life of the marginalised population, such as women in the fields of Science, Technology, Engineering, and Management (STEM), people who belong to the Lesbian, Gay, Bisexual, Transgender, Intersex and Queer/Questioning (LGBTIQ) community, racial minorities or people with disabilities have been studied, and it has been proven that the online life of vulnerable individuals produces a significant impact on their self-definition (Chung, 2013; Altszyler et al., 2018; Tortoreto et al., 2019). Furthermore, according to Milne et al. (2016), Burnap et al. (2017) and Kitzie (2018), the social networking activities of a vulnerable individual play an essential role in shaping the personality of the individual and how they look at society.

Comments/posts on online social media have been analysed to find and stop the spread of negativity using methods such as hate speech detection (Schmidt and Wiegand, 2017), offensive language identification (Zampieri et al., 2019a) and abusive language detection (Lee et al., 2018). According to Davidson et al. (2019), technologies developed for the detection of abusive language do not consider the potential biases of the dataset that they are trained on. The systematic bias in the datasets causes abusive language detection to be biased and may discriminate against one group over another. This will have a





*Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media*, pages 41–53
Barcelona, Spain (Online), December 13, 2020.

negative impact on minorities. We should turn our work towards spreading positivity instead of curbing an individual's freedom of speech by removing negative comments.

Therefore, we turn our research focus towards hope speech. Hope is commonly associated with the promise, potential, support, reassurance, suggestions or inspiration provided to participants by their peers during periods of illness, stress, loneliness and depression (Snyder et al., 2002). Psychologists, sociologists and social workers in the Association of Hope have concluded that hope can also be a useful tool for saving people from suicide or harming themselves (Herrestad and Biong, 2010). The Hope Speech delivered by gay rights activist Harvey Milk on the steps of the San Francisco City Hall during a mass rally to celebrate California Gay Freedom Day on 25 June 1978 [1] inspired millions to demand rights for equality, diversity and inclusion (Milk, 1997). However, to the best of our knowledge, no prior work has explored hope speech for women in STEM, LGBTIQ individuals, racial minorities or people with disabilities in general.

Moreover, although people of various linguistic backgrounds are exposed to online social media language, English is still at the centre of ongoing trends in language technology research. Recently, some research studies have been conducted on high resourced languages, such as Arabic, German, Hindi and Italian. However, such studies usually use monolingual corpora and do not examine code-switched textual data. Code-switching is a phenomenon where the individual switches between two or more languages in a single utterance (Sciullo et al., 1986). We introduce a dataset for hope speech identification not only in English but also in under-resourced code-switched Tamil (ISO 639-3: tam) and Malayalam (ISO 639-3: mal) languages (Chakravarthi et al., 2019; Jose et al., 2020; Priyadharshini et al., 2020).

The key contributions of this paper can be summarised as follows:

- We propose to encourage hope speech rather than take away an individual's freedom of speech by detecting and removing a negative comment.
- We apply the schema to create a multilingual, hostility-diffusing hope speech dataset for equality, diversity and inclusion. This is a new large-scale dataset of English, Tamil (code-switched), and Malayalam (code-switched) YouTube comments with high-quality annotation of the target.
- We performed an experiment on Hope Speech dataset for Equality, Diversity and Inclusion (HopeEDI) using different state-of-the-art machine learning models to create benchmark systems.

## 2 Related Works

When it comes to crawling social media data, there are many works on YouTube mining (Marrese-Taylor et al., 2017; Muralidhar et al., 2018), mainly focused on exploiting user comments. Krishna et al. (2013) did an opinion mining and trend analysis on YouTube comments. The researchers made an analysis of the sentiments to identify their trends, seasonality, and forecasts, and it was found that user sentiments are well correlated with the influence of real-world events. Severyn et al. (2014) did a systematic study on opinion mining targeting YouTube comments. The authors developed a comment corpus containing 35K manually labelled data for modelling the opinion polarity of the comments based on tree kernel models. Chakravarthi et al. (2020a) and Chakravarthi et al. (2020b) collected comments from YouTube and created a manually annotated corpus for the sentiment analysis of under-resourced Tamil and Malayalam languages.

Methods to mitigate gender bias in natural language processing (NLP) have been extensively studied for the English language (Sun et al., 2019). Some studies have investigated gender bias beyond the English language using machine translation to French (Vanmassenhove et al., 2018) and other languages (Prates et al., 2020). Tatman (2017) studied the gender and dialect bias in automatically generated captions from YouTube. Technologies for abusive language (Waseem et al., 2017; Clarke and Grieve, 2017), hate speech (Schmidt and Wiegand, 2017; Ousidhoum et al., 2019) and offensive language detection (Nogueira dos Santos et al., 2018; Zampieri et al., 2019b; Sigurbergsson and Derczynski, 2020) are being developed and applied without considering the potential biases (Davidson et al., 2019; Wiegand

[1] http://www.terpconnect.umd.edu/~jklumpp/ARD/MilkSpeech.pdf

et al., 2019; Xia et al., 2020). However, current gender debiasing methods in NLP are not sufficient to debias other issues related to EDI in end-to-end systems of many language technology applications, which causes unrest and escalates the issues with EDI, as well as leading to more inequality on digital platforms (Robinson et al., 2020).

Counter-narratives (i.e. informed textual responses) is another strategy, which has received the attention of researchers recently (Chung et al., 2019; Tekiroğlu et al., 2020). A counter-narrative approach was proposed to weigh the right to freedom of speech and avoid over-blocking. Mathew et al. (2019) created and released a dataset for counterspeech using comments from YouTube. However, the core idea to directly intervene with textual responses escalates hostility even though it is advantageous to the writer to understand why their comment/post has been deleted or blocked and then favourably change the discourse and attitudes of their comments. So we turn our research to finding positive information such as hope and encouraging such activities.

Recently, a work by Palakodety et al. (2020a) and Palakodety et al. (2020b) analysed how to use hope speech from a social media text to diffuse tension between two nuclear power nations (India and Pakistan) and support minority Rohingyas refugees. However, the author's definition of hope is just defined to diffuse tensions and violence. It does not take other perspectives of hope and EDI. The authors did not give more information such as the inter-annotator agreement, diversity in annotators and the details of the dataset. The dataset is not publicly available for research. It was created in English, Hindi and other languages related known to the Rohingyas. Our work differs from the previous works in that we define hope speech for EDI, and we introduce a dataset for English, Tamil and Malayalam on EDI of it. To the best of our knowledge, this is the first work to create a dataset for EDI in Tamil and Malayalam, which are under-resourced languages.

## 3 Hope Speech

Hope is considered significant for the well-being, recuperation and restoration of human life by health professionals. Hope can be defined as an optimistic state of mind that depends on a desire for positive results regarding the occasions and conditions of one's life or the world at large, and it is also present- and future-oriented (Snyder et al., 2002). Hope can also come from inspirational talk about how people face difficult situations and survive them. Hope speech engenders optimism and resilience that positively influences many aspects of life, including work (Youssef and Luthans, 2007), college (Chang, 1998) and other aspects that make us vulnerable (Cover, 2013). We define hope speech for our problem as "YouTube comments/posts that offer support, reassurance, suggestions, inspiration and insight".

Hope speech reflects the belief that one can discover pathways to their desired objectives and become roused to utilise those pathways. Our work aims to change the prevalent way of thinking by moving away from a preoccupation with discrimination, loneliness or the worst things in life to building the confidence, support and good qualities based on comments by individuals. Thus, we have provided instructions to annotators that if a comment/post meets the following conditions, then it should be annotated as hope speech.

- The comment contains inspiration provided to participants by their peers and others, offers support, reassurance, suggestions and insight
- The comment promotes well-being and satisfaction (past), joy, sensual pleasures and happiness (present).
- The comment triggers constructive cognition about the future – optimism, hope and faith.
- The comment contains an expression of love, courage, interpersonal skill, aesthetic sensibility, perseverance, forgiveness, tolerance, future-mindedness, praise for talents and wisdom.
- The comment encourages compliance with COVID-19 health guidelines.
- The comment promotes the values of equality, diversity and inclusion.

- The comment brings out a survival story of gay, lesbian or transgender individuals, women in science, or a COVID-19 survivor.
- The comment talks about fairness in the industry. (e.g., [I do not think banning all apps is right, we should ban only the apps which are not safe])
- Comments explicitly talking about a hopeful future. (e.g., [We will survive these things])
- Comments that explicitly talk about and say no to division in any form.
- The comment expresses positive peace-seeking intent (e.g., [We want peace]).

Non-hope speech includes comments that do not bring positivity, such as the following:

- The comment uses racially, ethnically, sexual or nationally motivated slurs.
- The comment produces hate toward a minority.
- The comment is very prejudiced and attacks people without thinking about the consequences.
- The comments do not inspire hope in the reader's mind.

Non-hope speech is different from hate speech. Some examples are shown below.

- **"How is that the same thing???"** This is non-hope speech but it is not hate speech either.
- **"Society says don't assume but they assume to anyways"** This is non-hope speech but it is not hate speech either.

Hate speech or offensive language detection dataset is not available for code-mixed Tamil and code-mixed Malayalam (Banerjee et al., 2020), and it does not take into account LGBTIQ, women in STEM and other minorities. Thus, we cannot use existing hate speech or offensive language detection datasets to detect hope or non-hope for EDI of minorities.

## 4 Dataset Construction

We focused on collecting data from the social media comments on YouTube [2], which is the most widely used platform in the world to express an opinion about a particular video. We avoided taking comments from personal coming out stories of LGBTIQ people as it had references to personal details, we manually removed the videos for personal coming out stories. For English, we collected data on recent topics of EDI, including women in STEM, LGBTIQ issues, COVID-19, Black Lives Matters, United Kingdom (UK) versus China, United States of America (USA) versus China and Australia versus China from YouTube video comments. The data was collected from videos of people from English-speaking countries, such as Australia, Canada, the Republic of Ireland, United Kingdom, the United States of America and New Zealand.

For Tamil and Malayalam, we collected data from India on the recent topics regarding LGBTIQ issues, COVID-19, women in STEM, the Indo-China war and Dravidian affairs. India is a multilingual and a multi-racial country. Linguistically, India can be divided into three major language families, namely Dravidian, Indo-Aryan and Tibeto-Burman languages (Chakravarthi et al., 2019; Chakravarthi et al., 2020c; Hande et al., 2020; Chakravarthi, 2020). The recent dispute on the Indo-China border has triggered racism on the internet towards people with Mongoloid features even though they are Indians from the North-Eastern states. Similarly, the National Education Policy, which advocates for the introduction of Sanskrit or Hindi has escalated issues regarding the linguistic autonomy of Dravidian languages in the state of Tamil Nadu. We used the YouTube comment scraper [3] to collect comments. We collected data on the above topics from November 2019 to June 2020 . We believe that our dataset will diffuse hostility and inspire hope. Our dataset is produced as a multilingual resource to allow cross-lingual studies and approaches. In particular, it contains hope speech in English, Tamil and Malayalam.

[2] https://www.youtube.com/

[3] https://github.com/philbot9/youtube-comment-scraper

### 4.1 Code-Mixing

Code-mixing is a phenomenon where the speaker uses two or more languages in a single utterance. It is prominent in multilingual speakers' social media discourse. Traditionally code-mixing has been associated with inadequate or informal knowledge of the language. However, research has shown that it is frequent in user-generated social media contents. For a multilingual country like India, code-mixing is quite frequent (Barman et al., 2014; Bali et al., 2014; Gupta et al., 2018). As our data comes from YouTube, our Tamil and Malayalam dataset is code-mixed. We have come across all the three types of code-mixing, such as tag, inter-sentential and intra-sentential in our corpus. Our corpus also has code-mixing using Latin script and native script.

### 4.2 Ethical Concerns

Social media data is highly sensitive, and even more so when it is related to the minority population, such as the LGBTIQ community or women. We have taken full consideration to minimise the risk associated with individual identity in the data by removing personal information from dataset, such as names but not celebrity names. However, to study EDI, we needed to keep information relating to the following characteristics; racial, gender, sexual orientation, ethnic origin and philosophical beliefs. Annotators were only shown anonymised posts and agreed to make no attempts to contact the comment creator. The dataset will only be made available for research purpose to the researcher who agree to follow ethical guidelines.

| Language | | English | Tamil | Malayalam |
|---|---|---|---|---|
| Gender | Male | 4 | 2 | 2 |
| | Female | 5 | 3 | 5 |
| | Non-binary | 2 | 1 | 0 |
| Higher Education | Undergraduate | 1 | 0 | 0 |
| | Graduate | 4 | 4 | 5 |
| | Postgraduate | 6 | 2 | 2 |
| Nationality | | Ireland, UK, USA, Australia | India, Sri Lanka | India |
| Total | | 11 | 6 | 7 |

Table 1: Annotators

### 4.3 Annotation Setup

After the data collection phase, we cleaned the data using *Langdetect*[4] to identify the language of the comments and removed comments that were not in the specified languages. However, there were unintended comments of other languages in the cleaned corpus of the Tamil and Malayalam comments due to code-mixing at different levels. Finally, we identified three classes, two of which are hope- and not-hope based on our definition from Section 3, while the last (Other languages) were introduced to account for comments that were not in the required language. These specific sets of classes were selected because they provided an adequate level of generalisation for characterising the comments of the EDI hope speech dataset.

### 4.4 Annotators

We created Google forms to collect annotations from annotators. Each form contained a maximum of 100 comments, and each page contained a maximum of 10 comments to maintain the quality of annotation. We collected information on the gender, educational background and the medium of schooling of the annotator to know the diversity of the annotator and avoid bias. The annotators were warned that comments might have offensive language and abusive text. The annotator was given the choice to stop annotating if they found the comments to be too disturbing or something that they could not handle.

[4] https://pypi.org/project/langdetect/

We educated annotators by providing them with YouTube videos on EDI [5] [6] [7] [8]. A minimum of three annotators annotated each form. As a warm-up procedure, after the first form containing 100 comments were annotated by annotators, the results were checked manually. This scheme was utilised to refine their understanding of the assignment and to improve the understanding of EDI. A few annotators dropped out after the initial stage of annotating their first form, and those annotations were discarded. The annotators were asked to watch the EDI videos again and reread the annotation guidelines. From Table 1, we can see the statistics of annotators. For English language comments, annotators were from Australia, the Republic of Ireland, the United Kingdom and the United States of America. For Tamil, we were able to get annotations from both people from the state of Tamil Nadu of India and from Sri Lanka. Most of the annotators were graduate or post-graduate students.

### 4.5 Inter-Annotator Agreement

To aggregate the hope speech annotations from multiple annotators, we opted for the majority, the comments that did not have a majority in the first round were collected, and a separate Google form was created to annotate them by new annotators. We calculated the inter-annotator agreement after the final round of annotation. We report inter-annotator agreement using the Krippendorff's alpha for assessing the clarity of the annotation. The Krippendorff's alpha is a statistical measure of agreement among annotators to answer how much the resulting data can be relied upon to represent real data (Krippendorff, 1970). Although **Krippendorff's alpha** $(\alpha)$ is computationally complex, it is more relevant to our case as more than two annotators annotated the comments, and the same annotators did not annotate all the sentences. It is not affected by missing data, takes into account varying the sample sizes, categories, the numbers of raters and can also be employed for any measurement levels, such as nominal, ordinal, interval and ratio. We used *nltk*[9] for calculating Krippendorff's alpha $(\alpha)$ (Krippendorff, 2011). Our annotations produced an agreement of 0.63, 0.76, and 0.85 using nominal metric for English, Tamil and Malayalam respectively.

| Language pair | English | Tamil | Malayalam |
|---|---|---|---|
| Number of Words | 522,717 | 191,242 | 122,917 |
| Vocabulary Size | 29,383 | 46,237 | 40,893 |
| Number of Comments/Posts | 28,451 | 20,198 | 10,705 |
| Number of Sentences | 46,974 | 22935 | 13,643 |
| Average number of words per sentences | 18 | 9 | 11 |
| Average number of sentences per post | 1 | 1 | 1 |

Table 2: Corpus statistic

| Class | English | Tamil | Malayalam |
|---|---|---|---|
| Hope | 2,484 | 7,899 | 2,052 |
| Not Hope | 25,940 | 9,816 | 7,765 |
| Other lang | 27 | 2,483 | 888 |
| Total | 28,451 | 20,198 | 10,705 |

Table 3: Classwise Data Distribution

[5] https://www.youtube.com/watch?v=C-uyB5I6WnQ&t=6s
[6] https://www.youtube.com/watch?v=UcuS5glhNto
[7] https://www.youtube.com/watch?v=hNeR4bBUj68
[8] https://www.youtube.com/watch?v=LqP6iU3g2eE
[9] https://www.nltk.org/

| | English | Tamil | Malayalam |
|---|---|---|---|
| Training | 22,762 | 16,160 | 8564 |
| Development | 2,843 | 2,018 | 1070 |
| Test | 2,846 | 2,020 | 1071 |
| Total | 28,451 | 20,198 | 10,705 |

Table 4: Train-Development-Test Data Distribution

### 4.6 Corpus Statistics

In total, our dataset contains 59,354 comments from YouTube videos, where 28,451 comments in English, 20,198 comments in Tamil, and 10,705 comments in Malayalam. Table 2 shows the distribution of our dataset. We used *nltk* tool to tokenise words and sentences in the comments to calculate corpus statistics. As shown, the vocabulary for Tamil and Malayalam is high due to the different types of code-mixing.

Table 3 presents the distribution of the annotated dataset by label. The dataset is skewed, with almost the majority of the comments being labelled as not hope (NOT). This is common for user-generated content on online platforms, and an automatic detection system needs to be able to handle imbalanced data in order to be truly useful. We have a considerable amount of "Other language" labels for Tamil and Malayalam; this is also due to high code-mixing phenomenon occurring in the comments of these languages. The fully annotated dataset was split into a train, development and test set. The training set contains 80%, the development set contains 10% and finally, the test set contains the remaining 10% of the data shown in Table 4.

### 4.7 Ambiguous Comments

We found some ambiguous comments during the process of annotation.

- **"Chanting Black Lives Matter is Racist"** This sentence from the English corpus was confusing. The annotators were as confused as we were about comments like these.
- **"God gave us a choice"** This sentence was interpreted by some as hope and others as not-hope.
- **Sri Lankan Tamilar history patti pesunga** – *Please speak about history of Tamil people in Sri Lanka.* Inter-sentential switch in Tamil corpus written using Latin script. The history of Tamil people in Sri Lanka is both hopeful and non-hopeful due to the recent civil war.
- **Bro helo app ku oru alternate appa solunga.** – *Bro tell me an alternate app for Helo app.* Intra-sentential and tag switch in Tamil corpus written using Latin script.

## 5 Benchmark Experiments

We reported our dataset using a wide range of standard classifiers on the unbalanced settings of the dataset. The experiment was applied on the token frequency-inverse document frequency (Tf-Idf) of tokens. We used sklearn [10]library to create baseline classifiers. For the multinomial Naive Bayes, we set alpha = 0.7. We used a grid search for the k-nearest neighbors (KNN), support vector machine (SVM), decision tree and logistic regression. More details about the parameters of the classifier will be published in the code.

Our models were trained on the training dataset; the development set was used to fine-tune the model, and it was evaluated by predicting the labels for the held-out test set, as shown in Table 4. To report the performance of the classification, we used a macro-averaged F-score, calculated using macro-averaged precision and recall. The motivation behind such a choice is due to the imbalanced class distribution,

[10]https://scikit-learn.org/stable/

| Classifier | Hope Speech | Not-Hope Speech | Other language | Macro Avg | Weighted Avg |
|---|---|---|---|---|---|
| Support | 250 | 2,593 | 3 | | |
| **Precision** | | | | | |
| SVM | 0.00 | 0.91 | 0.00 | 0.30 | 0.83 |
| MNB | 0.14 | 0.91 | 0.00 | 0.35 | 0.84 |
| KNN | 0.63 | 0.92 | 0.00 | 0.52 | 0.90 |
| DT | 0.46 | 0.94 | 0.00 | 0.47 | 0.90 |
| LR | 0.33 | 0.96 | 0.00 | 0.43 | 0.90 |
| **Recall** | | | | | |
| SVM | 0.00 | 1.00 | 0.00 | 0.33 | 0.83 |
| MNB | 0.00 | 1.00 | 0.00 | 0.33 | 0.91 |
| KNN | 0.14 | 0.99 | 0.00 | 0.38 | 0.92 |
| DT | 0.39 | 0.96 | 0.00 | 0.45 | 0.90 |
| LR | 0.59 | 0.88 | 0.00 | 0.49 | 0.86 |
| **F-Score** | | | | | |
| SVM | 0.00 | 0.95 | 0.00 | 0.32 | 0.87 |
| MNB | 0.01 | 0.95 | 0.00 | 0.31 | 0.87 |
| KNN | 0.23 | 0.96 | 0.00 | 0.40 | 0.89 |
| DT | 0.42 | 0.95 | 0.00 | 0.46 | 0.90 |
| LR | 0.43 | 0.92 | 0.00 | 0.45 | 0.87 |

Table 5: Precision, Recall, and F-score for English

| Classifier | Hope Speech | Not-Hope Speech | Other language | Macro Avg | Weighted Avg |
|---|---|---|---|---|---|
| Support | 815 | 946 | 259 | | |
| **Precision** | | | | | |
| SVM | 0.00 | 0.47 | 0.00 | 0.16 | 0.22 |
| MNB | 0.58 | 0.57 | 0.74 | 0.63 | 0.60 |
| KNN | 0.48 | 0.55 | 0.55 | 0.53 | 0.52 |
| DT | 0.52 | 0.57 | 0.52 | 0.53 | 0.54 |
| LR | 0.59 | 0.59 | 0.47 | 0.55 | 0.58 |
| **Recall** | | | | | |
| SVM | 0.00 | 1.00 | 0.00 | 0.33 | 0.47 |
| MNB | 0.42 | 0.81 | 0.25 | 0.49 | 0.58 |
| KNN | 0.35 | 0.72 | 0.38 | 0.48 | 0.53 |
| DT | 0.40 | 0.71 | 0.41 | 0.51 | 0.55 |
| LR | 0.37 | 0.73 | 0.64 | 0.58 | 0.57 |
| **F-Score** | | | | | |
| SVM | 0.00 | 0.64 | 0.00 | 0.21 | 0.30 |
| MNB | 0.49 | 0.67 | 0.37 | 0.51 | 0.56 |
| KNN | 0.41 | 0.62 | 0.45 | 0.49 | 0.51 |
| DT | 0.45 | 0.63 | 0.46 | 0.51 | 0.53 |
| LR | 0.46 | 0.65 | 0.55 | 0.55 | 0.56 |

Table 6: Precision, Recall, and F-score for Tamil

which makes well-known measures such as accuracy and the micro-average F-score not well representative of the performance. Since the performance of all classes is of interest, we also reported the precision, recall and the weighted F-score of the individual classes. Table 5, Table 6 and Table 7 reports the precision, recall and F-score results of the test set of HopeEDI using baselines classifiers, alongside support from test data.

As shown, all the models performed poorly due to a class imbalance problem. The SVM classifier achieved the lowest performance on the HopeEDI dataset with a macro-average F-Score of 0.32, 0.21 and 0.28 for English, Tamil and Malayalam respectively. The decision tree had a higher macro F-Score for English and Malayalam while Tamil performed well in the logistic regression. We used language identification to remove the non-intended language comments from our dataset. However, there were some comments that were annotated by annotators as "Other language". This caused another imbalance in our dataset. Most of the macro scores were less for English due to the "Other language" label; this could be avoided for English by merely removing those comments in the dataset. However, for Tamil and Malayalam, this label was necessary as the comments in these languages were code-mixed and written using a non-native script (Latin script). For the Tamil language, the data distribution was somewhat

| Classifier | Hope Speech | Not-Hope Speech | Other language | Macro Avg | Weighted Avg |
|---|---|---|---|---|---|
| Support | 194 | 776 | 101 | | |
| **Precision** | | | | | |
| SVM | 0.00 | 0.72 | 0.00 | 0.24 | 0.52 |
| MNB | 0.78 | 0.76 | 0.91 | 0.81 | 0.78 |
| KNN | 0.39 | 0.77 | 0.79 | 0.65 | 0.71 |
| DT | 0.51 | 0.81 | 0.52 | 0.61 | 0.73 |
| LR | 0.46 | 0.79 | 0.45 | 0.57 | 0.70 |
| **Recall** | | | | | |
| SVM | 0.00 | 1.00 | 0.00 | 0.33 | 0.72 |
| MNB | 0.16 | 1.00 | 0.10 | 0.42 | 0.76 |
| KNN | 0.12 | 0.96 | 0.37 | 0.48 | 0.75 |
| DT | 0.27 | 0.92 | 0.40 | 0.53 | 0.76 |
| LR | 0.25 | 0.89 | 0.39 | 0.51 | 0.73 |
| **F-Score** | | | | | |
| SVM | 0.00 | 0.84 | 0.00 | 0.28 | 0.61 |
| MNB | 0.26 | 0.86 | 0.18 | 0.44 | 0.69 |
| KNN | 0.19 | 0.86 | 0.50 | 0.51 | 0.70 |
| DT | 0.36 | 0.86 | 0.45 | 0.56 | 0.73 |
| LR | 0.33 | 0.84 | 0.41 | 0.53 | 0.70 |

Table 7: Precision, Recall, and F-score for Malayalam

balanced between hope and non-hope classes.

In order to evaluate the effectiveness of our dataset, we conducted experiments using machine learning algorithms. We believe the HopeEDI dataset, with its novel method of data collection and annotation, shall revolutionise research in language technology in the future broaden the horizon for further research on positivity.

## 6 Conclusion

As online content increases massively, it is necessary to encourage positivity such as in the form of hope speech in online forums to induce compassion and acceptable social behaviour. In this paper, we presented the largest manually annotated dataset of hope speech detection in English, Tamil and Malayalam, consisting of 28,451, 20,198 and 10,705 comments, respectively. We believe that this dataset will facilitate future research on encouraging positivity. We aim to promote research in hope speech and to encourage positive content in online social media for equality, diversity and inclusion. In the future, we plan to extend the study by introducing a larger dataset with further fine-grained classification and content analysis.

## 7 Acknowledgments

The author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2), co-funded by the European Regional Development Fund as well as by the EU H2020 programme under grant agreement 825182 (Prêt-à-LLOD), and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) for his postdoctoral period at National University of Ireland Galway.

## References

Edgar Altszyler, Ariel J. Berenstein, David Milne, Rafael A. Calvo, and Diego Fernandez Slezak. 2018. Using contextual information for automatic triage of posts in a peer-support forum. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 57–68, New Orleans, LA, June. Association for Computational Linguistics.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar, October. Association for Computational Linguistics.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2020. Comparison of pretrained embeddings to identify hate speech in Indian code-mixed text. In *2nd IEEE International Conference on Advances in Computing, Communication Control and Networking –ICACCCN (ICAC3N-20)*.

Utsab Barman, Joachim Wagner, Grzegorz Chrupała, and Jennifer Foster. 2014. DCU-UVT: Word-level language classification with code-mixed data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 127–132, Doha, Qatar, October. Association for Computational Linguistics.

Pete Burnap, Gualtiero Colombo, Rosie Amery, Andrei Hodorog, and Jonathan Scourfield. 2017. Multi-class machine classification of suicide-related communication on twitter. *Online Social Networks and Media*, 2:32 – 44.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland, 19 August. European Association for Machine Translation.

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

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France, May. European Language Resources association.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E.O'Connor, and John P McCrae. 2020c. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects*, Barcelona, Spain, December.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Edward C. Chang. 1998. Hope, problem-solving ability, and coping in a college student population: Some implications for theory and practice. *Journal of Clinical Psychology*, 54(7):953–962.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July. Association for Computational Linguistics.

Jae Eun Chung. 2013. Social networking in online support groups for health: How online social networking benefits patients. *Journal of Health Communication*, 19(6):639–659, April.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Rob Cover. 2013. Queer youth resilience: Critiquing the discourse of hope and hopelessness in lgbt suicide representation. *M/C Journal*, 16(5).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August. Association for Computational Linguistics.

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

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Uncovering code-mixed challenges: A framework for linguistically driven question generation and neural based question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 119–130, Brussels, Belgium, October. Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, Barcelona, Spain, December.

Henning Herrestad and Stian Biong. 2010. Relational hopes: A study of the lived experience of hope in some patients hospitalized for intentional self-harm. *International Journal of Qualitative Studies on Health and Well-being*, 5(1):4651. PMID: 20640026.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*.

Vanessa Kitzie. 2018. "i pretended to be a boy on the internet": Navigating affordances and constraints of social networking sites and search engines for lgbtq+ identity work. *First Monday*, 23(7), Jul.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Amar Krishna, Joseph Zambreno, and Sandeep Krishnan. 2013. Polarity Trend Analysis of Public Sentiment on YouTube. In *Proceedings of the 19th International Conference on Management of Data*, COMAD '13, page 125–128, Mumbai, Maharashtra, IND. Computer Society of India.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. 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. Association for Computational Linguistics.

Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2017. Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 102–111, Copenhagen, Denmark, September. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):369–380, Jul.

Harvey Milk. 1997. The hope speech. *We are everywhere: A historical sourcebook of gay and lesbian politics*, pages 51–53.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, San Diego, CA, USA, June. Association for Computational Linguistics.

Skanda Muralidhar, Laurent Nguyen, and Daniel Gatica-Perez. 2018. Words worth: Verbal content and hirability impressions in YouTube video resumes. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 322–327, Brussels, Belgium, October. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia, July. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. 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. Association for Computational Linguistics.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020a. Hope speech detection: A computational analysis of the voice of peace. In *Proceedings of the 24th European Conference on Artificial Intelligence - ECAI 2020*.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020b. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 454–462.

Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381, May.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named entity recognition for code-mixed Indian corpus using meta embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*.

Laura Robinson, Jeremy Schulz, Grant Blank, Massimo Ragnedda, Hiroshi Ono, Bernie Hogan, Gustavo S. Mesch, Shelia R. Cotten, Susan B. Kretchmer, Timothy M. Hale, Tomasz Drabowicz, Pu Yan, Barry Wellman, Molly-Gloria Harper, Anabel Quan-Haase, Hopeton S. Dunn, Antonio A. Casilli, Paola Tubaro, Rod Carvath, Wenhong Chen, Julie B. Wiest, Matías Dodel, Michael J. Stern, Christopher Ball, Kuo-Ting Huang, and Aneka Khilnani. 2020. Digital inequalities 2.0: Legacy inequalities in the information age. *First Monday*, 25(7), Jun.

Anna Schmidt and Michael Wiegand. 2017. 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. Association for Computational Linguistics.

Anne-Marie Di Sciullo, Pieter Muysken, and Rajendra Singh. 1986. Government and code-mixing. *Journal of Linguistics*, 22(1):1–24.

Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2014. Opinion mining on YouTube. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1252–1261, Baltimore, Maryland, June. Association for Computational Linguistics.

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

Charles R Snyder, Kevin L Rand, and David R Sigmon. 2002. Hope theory: A member of the positive psychology family.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online, July. Association for Computational Linguistics.

Giuliano Tortoreto, Evgeny Stepanov, Alessandra Cervone, Mateusz Dubiel, and Giuseppe Riccardi. 2019. Affective behaviour analysis of on-line user interactions: Are on-line support groups more therapeutic than twitter? In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 79–88, Florence, Italy, August. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zijian Wang and David Jurgens. 2018. 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. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online, July. Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark, September. Association for Computational Linguistics.

Carolyn M. Youssef and Fred Luthans. 2007. Positive organizational behavior in the workplace: The impact of hope, optimism, and resilience. *Journal of Management*, 33(5):774–800.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. 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. Association for Computational Linguistics.