Stop the Hate, Spread the Hope: An Ensemble Model for Hope Speech Detection in English and Dravidian Languages

DEEPAWALI SHARMA

Department of Computer Science, Banaras Hindu University Faculty of Science, Varanasi, Uttar Pradesh, India, deepawali21@bhu.ac.in

School of Computer Science Engineering and Technology (SCSET), Bennett University, Noida, Uttar Pradesh, India, deepawali21@bhu.ac.in

VEDIKA GUPTA

Jindal Global Business School, OP Jindal Global University, Sonipat, Haryana, India, vedika.nit@gmail.com

VIVEK KUMAR SINGH

Department of Computer Science, Banaras Hindu University Faculty of Science, Varanasi, Uttar Pradesh, India, vivekks12@gmail.com

Department of Computer Science, University of Delhi, New Delhi, Delhi, India, vivekks12@gmail.com

BHARATHI RAJA CHAKRAVARTHI

School of Computer Science, National University of Ireland Galway, Galway, Galway, Ireland, bharathiraja.asokachakravarthi@nuigalway.ie

Unit for Linguistic Data, Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Galway, Galway, Ireland, bharathiraja.asokachakravarthi@nuigalway.ie

The rise of social media has led to vast amounts of user-generated content, with emotions ranging from joy to anger. Negative comments often target individuals, communities, or brands, prompting successful efforts to detect harmful speech such as hate speech, cyberbullying, and abuse. Recently, another type of speech referred to as 'Hope Speech' has gained attention from the research community. Hope speech consists of positive affirmations or words of reassurance, encouragement, consolation or motivation offered to the affected individual/ community during the lean periods of life. However, there has been relatively less research focused on the detection of hope speech, more particularly in low-resource languages. This paper, therefore, attempts to develop an ensemble model for detecting hope speech in some low-resource languages. Data for four different languages, namely English, Kannada, Malayalam and Tamil are obtained and experimented with different deep learning-based models. An ensemble model is proposed to combine the advantages of the better performing models. Experimental results demonstrate the superior performance of the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) model compared to individual models based on data from all four languages (weighted average F1-score for English is 0.93; for Kannada is 0.74; for Malayalam is 0.82; and for Tamil is 0.60). Thus, the proposed ensemble model proves to be a suitable approach for hope speech detection in the given low resource languages.

Keywords: CNN, Deep Learning, GloVe embedding, Hate Speech, Hope Speech, Low Resource Language, LSTM, mBERT, XLM-RoBERTa.

Disclaimer—*The given study involves racial slurs, aggravated, and the use of harmful words targeted especially towards women, LGBTQ+, people of color, and other communities. However, given the nature of the study, they cannot be overlooked. The paper is solely for research purposes and doesn't support hate and aggressive speech in any manner towards anyone.*

ACM 2375-4699/2025/02-ART
http://dx.doi.org/10.1145/3716383

1 Introduction

The online social media penetration across the world has increased significantly during recent period, reaching to more than 4.62 billion active social media users as in Jan 2022[1] (which is 58.4% of the world population). Social media platforms allow anyone and everyone to create and post content in different online social platforms. This has led to creation of huge volume of user generated content on the Internet. A significant part of this data is unstructured in nature and a large volume out of that comprises of textual data. Several online social media platforms like Facebook, Blog, and Twitter allow posting textual comments/ opinions/ reviews. However, the freedom of posting content on these platforms sometimes also result in posting of some undesired or offensive content. Advances in Natural Language Processing (NLP) combined with computing power have attracted researchers across the world to design automated techniques and approaches for analysing the text available on these platforms.

Although majority of the Internet users resort to using legitimate choice of words while sharing their viewpoints, still some users violate these social norms and use inadmissible phrases or speech. Such kind of content leads to spreading negativity through the social media platforms. It is observed that vulnerable communities such as Lesbian, Gay, Bisexual, Transgender, Intersex and Questioning (LGBTIQ), people of color, and women usually face the wrath of such negativity [5][35]. Therefore, researchers are now working towards design of automated techniques and approaches for dealing with such content. The detection of offensive, hateful, defaming, and unwanted comments on gender, religion, race, or ethnicity on social media has been enormously studied in literature [1][6][11][18][23]. Existing studies have focussed not only on detecting the hate speech, and cyberbullying but also contributed towards intuitive ways to handle these comments [8][25-26][36-37].

While on one side, social media platform is being misused for hate speech, on the other side, it has also emerged as a wonderful and accessible platform to spread positivity and seek hope. It has been witnessed that people suffering from anxiety or depression, search for a source of encouragement. Most of the people find this source in their known circles- friends or relatives, whereas some people prefer to seek hope or positivity from people or communities on social media or virtual platforms. This type of speech on social media, often referred to as 'Hope Speech', is gaining familiarity on the Internet in recent times. Hope speech is a positive affirmation or words of reassurance, encouragement, consolation or motivation provided to the affected individual/ community during the lean periods of life [12]. Those comments that offer encouragement, inspiration, and support are considered hope speech, but those comments that demotivate, and lower the subject's value are considered not hope speech. Hope speech spreads positivism. Therefore, detection of hope speech could help promoting spread of such content, which can be beneficial in a number of ways.

A number of studies have been conducted on the detection of hope speech. Such studies have mainly explored applicability of deep learning models on English language and also on two Indian languages- Malayalam and Tamil [4][6][7][2]. Another set of studies have also explored hope speech detection in Kannada language [23][24]. Kannada, Malayalam and Tamil languages belong to Dravidian family of languages of India. Another recent study explored hope speech detection in Spanish language text from social media [31] [40]. There are, however, limited number of studies on hope speech detection as compared to hate speech. Further, the low resource languages are not much explored with respect to the task of hope speech detection. The present work, therefore, attempts to explore the suitability of an ensemble approach involving transformer-based models (mBERT, XLM-RoBERTa) and deep learning models (CNN, LSTM) for hope speech detection in four selected languages, namely English, Kannada, Malayalam and Tamil. Out of the above-mentioned four languages, only the English language is a language with rich lexical resources for Natural Language Processing. The other languages are called low or scarce resource languages because these languages have a smaller number of lexical resources that have been developed or are in developing stages [38]. Due to the lack of resources, the comments, perspectives, and opinions

[1] https://datareportal.com/reports/digital-2022-global-overview-report

generated online in these languages are not sufficiently tapped. Therefore, this paper attempts to perform hope speech detection for three Dravidian languages (Kannada, Malayalam, and Tamil) along with the English language through a suitable algorithmic approach.

1.1 Research Questions

The paper attempts to address the following research questions:

RQ1: Which computational architectures can be suitably employed to detect hope speech from data comprising YouTube comments, particularly in low-resource languages?

RQ2: Whether an ensemble architecture of better-performing models can further improve the performance on the task?

1.2 Major Contributions

The dataset provided by Language Technology for Equality, Diversity and Inclusion (LT-EDI-ACL 2022) is used for the experimental purpose. The comments are classified as 'hope speech' or 'not hope speech' for EDI. Deep learning models (CNN, LSTM) using GloVe embedding, transformer-based models (mBERT, XLM-RoBERTa) and Ensemble (LSTM, mBERT, XLM-RoBERTa) models are used to detect hope speech.

More specifically, the following tasks are performed, which contribute to the existing research work in the area:

- Various deep learning models (CNN, LSTM) using GloVe embedding and transformer-based models (mBERT, XLM-RoBERTa) for different languages (English, Kannada, Malayalam and Tamil) are explored for suitability for hope speech detection task.
- An ensemble model (called Ensemble (LSTM, mBERT, XLM-RoBERTa)) is designed with the aim of further improving the performance on the task by combining advantages of better performing deep learning and transformer-based models for each language (English, Kannada, Malayalam and Tamil).
- The results obtained are compared with some previous research work and is found to achieve superior performance.

The rest of the paper is organized as follows: Section 2 discusses the related work on hope speech detection in multilingual as well as code-mixed language datasets. Section 3 describes the dataset used and its dimensions. Section 4 presents the architecture of the various models implemented, including the technical architecture of the proposed ensemble model. Results are presented in Section 5 along with a comparison with previous studies and a brief discussion on Error Analysis. Section 6 provides the analysis and details of hyperparameters of the proposed model. The paper concludes in Section 7 with a summary of the results and major findings.

2 Related Work

The comments that spread hate, and outrage do no good for an individual or society. Therefore, several studies have been conducted to detect and flag such hate speech. Some studies proposed a methodology to identify hate speech forms and discover a set of patterns and a broader understanding of online hate speech [15]. Similarly, for tweet classification as racist, sexist or neither, experimentation has been done with multiple deep-learning architectures [17]. Some of them describe pilot classification experiments to classify anti-Semitic speech [19]. As compared to the detection of hate speech, there are very few existing studies on the automated detection of hope speech. The Hope for Equality, Diversity, and Inclusion (HopeEDI) dataset was used for the detection of hope speech in English and Dravidian languages (Kannada, Malayalam, and Tamil) in one of first research works on the topic [39]. This dataset continues to remain the most important dataset for the task. This section discusses some of the prominent studies that have experimented on the HopeEDI dataset using various machine learning, deep learning, hybrid, and transformer-based models. Authors in [32] used XGBoost, Random Forest (RF), Multinomial Naïve Bayes (NB), and Logistic regression (LR) to detect whether a comment is either hope speech or not.

Nowadays, in addition to classical machine learning models, deep neural networks and hybrid models are also deployed, and these models improved performance considerably over traditional/classical models.

Some of the prominent studies that experimented with deep learning and hybrid models along with machine learning use LR and Support Vector Machine (SVM) with Term Frequency-Inverse Document Frequency (TF-IDF). Several Deep learning models were also proposed, such as LSTM, Bidirectional-LSTM (Bi-LSTM) using different word embedding, 2-parallel CNN-LSTM, 3-parallel Bi-LSTM, etc. [2]. Similarly, different deep neural networks DNN, CNN, Bi-LSTM, and hybrid model LSTM-CNN were also used to detect hope speech [3]. Recently, transformer embeddings used with deep learning models have also proven to be effective [29]. The BiLSTM models (embedded with mBERT, XLM-RoBERTa and many other embeddings) were used for English, Tamil, and Malayalam [9]. The problem of hope speech detection has also been approached using character n-grams-based TF-IDF and Multilingual Representations for Indian Languages (MuRIL) text representations [4]. The authors have also compared different approaches, namely TF-IDF + LR, TF-IDF + SVM, MuRIL + LR and MuRIL + SVM for each language in their dataset.

Transformer-based models have also recently gained popularity in the field of NLP. With the attention mechanism, the transformer allows for the parallelization of input rather than processing one word at a time [33]. Various transformer-based models like mBERT, XLM-RoBERTa and many more are now regarded the state-of-the-art models [7]. Therefore, for detecting hope in comments, several transformer-based models namely mBERT, IndicBERT, and XLM-RoBERTa framework have been deployed and the labels were classified using the output obtained from the final layer of XLM-RoBERTa [6]. A transformer-based approach has been proposed for hope speech detection in four different languages (English, Tamil, Malayalam, and Kannada) [24].

Some authors also worked on code-mixed language to detect hope in comments. The code-mixed language contains more than one language in the same dataset. A Dual-Channel Language model (DC-LM) was proposed by fine-tuning a language model based on the transformer architecture on the code-mixed data and its translation in English and Google translation API [2] was used to translate the code-mixed KanHope to English [22].

The first hope speech detection dataset was created by Bharathi Raja Chakravarthi in the year 2020 [5]. The hope speech dataset for Equality, diversity and inclusion (HopeEDI) in different languages created by sourcing the social media comments, is now one of the most important datasets for the task [11]. Most of the studies on hope speech detection have therefore used this dataset. **Table 1** presents a summary of key results and limitations of existing studies hope speech detection, most of which have used the HopeEDI dataset.

Table 1: Tabular representation of some previous research work in the area

Approach	Author	Results				Limitations
		Language	Model	Embedding	Weighted F1 score	
	Gupta, V., Kumar, R., & Pamula, R. (2022, May) [32]	English	XGBoost Random Forest Multinomial NB Logistic Regression	- - - -	0.86 0.86 0.81 0.83	1. Detected hope speech only in the English language, although the dataset is available in Dravidian languages as well. 2. Implemented only classical machine learning models and did not experiment with deep learning and transformer-based models.
	Jha, V., Mishra, A., & Saumya, S. (2022, May) [3]	English	DNN CNN Bi-LSTM GRU	- - - -	0.89 0.88 0.88 0.88	1. Detected hope speech only in the English language, although the dataset is available in Dravidian

[2] https://translate.google.com/

Machine Learning/ Deep Learning-based Approach			LSTM-CNN	-	0.87	languages as well.
			LSTM-LSTM	-	0.89	
	Saumya, S., & Mishra, A. K. (2021, April) [2]	English	SVM	Tf-idf	0.85	1. No experimentation and evaluation for Kannada language.
			LR	Tf-idf	0.75	
			LSTM	GloVe	0.90	
			Bi-LSTM	GloVe	0.90	
			2 Layered CNN	Random	0.79	
			2-parallel CNN-LSTM	GloVe, Word2Vec	0.91	
		Tamil	SVM	Tf-idf	0.48	
			LR	Tf-idf	0.50	
			Bi-LSTM	Random	0.52	
			CNN	Word2Vec	0.55	
			2 layered CNN	Random	0.55	
			3-parallel Bi-LSTM	Word2Vec, Random	0.56	
		Malayalam	SVM	Tf-idf	0.65	
			LR	Tf-idf	0.72	
			CNN	Word2Vec	0.70	
			LSTM	Word2Vec	0.75	
			Bi-LSTM	Random	0.74	
			3-parallel Bi-LSTM	Word2Vec, Random	0.78	
	Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., & Chakravarthi, B. R. (2021) [9]	English	BiLSTM	bert -base – uncased	0.92	1. No direct experimentation with different models (BERT, XLM-RoBERTa, MuRIL). 2. No experimentation and evaluation for Kannada language.
		Tamil	BiLSTM	bert -base – uncased	0.56	
				mBERT-cased	0.52	
				xlm-roberta-base	0.57	
				MuRIL	0.54	
		Malayalam	BiLSTM	bert -base – uncased	0.85	
				mbert-cased	0.84	
				xlm-roberta-base	0.82	
				MuRIL	0.82	
Transformer based Approach	Arunima, S., Ramakrishnan, A., Balaji, A., & Thenmozhi, D. (2021, April) [6]	English	BERT	-	0.92	1. Only experimented with only one transformer-based model. 2. No experimentation and evaluation for Kannada language.
		Tamil	mBERT	-	0.46	
		Malayalam	mBERT	-	0.81	
	Singh, P., Kumar, P., & Bhattacharyya, P. (2021, April) [7]	English	mBERT	-	0.92	1. No experimentation and evaluation for Kannada language.
			XLM-RoBERTa	-	0.92	
			IndicBERT	-	0.91	
			MuRIL	-	0.91	
		Tamil	mBERT	-	0.57	
			XLM-RoBERTa	-	0.58	
			IndicBERT	-	0.54	
			MuRIL	-	0.56	
		Malayalam	mBERT	-	0.85	
			XLM-RoBERTa	-	0.86	
			IndicBERT	-	0.83	
			MuRIL	-	0.82	
	Vijayakumar, P., Prathyush, S., Aravind, P., Angel, S., Sivanaiah, R., Rajendram, S. M., & Mirnalinee, T. T. (2022, May) [24]	English	ALBERT	-	0.88	1. Only experimented with the BERT-based model (ALBERT). Other transformer-based models not implemented.
		Tamil	ALBERT	-	0.39	
		Malayalam	ALBERT	-	0.74	
		Kannada	ALBERT	-	0.75	
	Zhu, Y. (2022, May). [13]	English	LSTM	-	0.67	1. The transformer-based models and their ensemble can be further improved for superior performance.
			CNN	-	0.59	
			CNN+LSTM	-	0.70	
			BiLSTM	-	0.71	
			CNN+BiLSTM	-	0.75	

Ensemble Approach			LSTM+BiLSTM Ensemble(CNN, CNN+LSTM,BiLSTM	- -	0.80 0.88	
		Tamil	LSTM Ensemble(CNN, CNN+LSTM,BiLSTM	- -	0.30 0.41	
		Malayalam	LSTM CNN CNN+LSTM BiLSTM CNN+BiLSTM LSTM+BiLSTM Ensemble(CNN, CNN+LSTM,BiLSTM	- - - - - - -	0.64 0.55 0.66 0.64 0.66 0.70 0.72	
		Kannada	LSTM Ensemble(CNN, CNN+LSTM,BiLSTM	- -	0.57 0.72	
	Kumar, A., Saumya, S., & Roy, P. (2022, May) [23]	English	Ensemble model (TF-IDF (SVM,LR and Random Forest)	-	0.88	1. Only classical machine learning and an ensemble of machine learning models are implemented. Not experimented with deep learning models and transformer-based models.
		Tamil	Ensemble model (TF-IDF (SVM,LR and Random Forest)	-	0.38	
		Malayalam	Ensemble model (TF-IDF (SVM,LR and Random Forest)	-	0.74	
		Kannada	SVM Logistic Regression Random Forest Ensemble model (TF-IDF (SVM, LR, and Random Forest)	- - - -	0.75	

There are fewer studies that have used ensemble approaches involving advanced models to detect hope speech from natural language. Basically, the ensemble model combines the prediction of two or more baseline models and results in one final prediction [34]. As presented in Table 1, authors in [13] have developed an ensemble of different deep-learning baseline models like CNN, LSTM, and BiLSTM, while an ensemble of machine-learning models with the use of TF-IDF has also been proposed [23]. However, some of the recently proposed deep learning and transformer-based models have not been explored for the task. Similarly, an ensemble of such advanced models has also not been explored.

The field of hope speech detection in natural language presents some interesting open challenges. Some of them can be as follows:

i) Currently, the online social media data is being generated in regional or code-mixed languages. Processing such data is a challenging task due to the unavailability of suitable lexical resources in different low-resource languages. Therefore, suitable datasets for low resource languages are required.
ii) The current advances lack a generic or unified model that can accurately detect hope speech in multiple languages, including the low-resource languages. Therefore, a suitable model that works on multiple languages would be a good contribution in the area.
iii) Though ensemble approach can combine advantages of individual models for better accuracy, the ensemble of advanced models (deep learning and transformer-based models) has not been sufficiently explored for the task of hope speech detection. Therefore, such an exercise will be highly useful.
iv) Sometimes people use sarcastic sentences in their comments (such as comments where positive expressions are used but they mock or convey contempt). Such comments are often misclassified by the various

computational models. Therefore, effort is also needed in this direction to handle sarcasm in hope speech detection.

The present work attempts to address challenges (ii) and (iii). An attempt is made to work towards improving the performance of computational approach for hope speech detection by experimenting with advanced deep learning and transformer-based models. An ensemble of better-performing models for the task is also designed. Further, it is shown that the ensemble model can work suitably on data for multiple languages (the considered languages are English, Kannada, Malayalam, Tamil).

3 Dataset Details

3.1 Dataset Collection

This study uses one of the most popular datasets on hope speech, the HopeEDI[3] dataset. This dataset was launched by LT-EDI in the year 2021 and is available online at: https://competitions.codalab.org/competitions/. The dataset contains YouTube comments in four different languages, namely English, Kannada, Malayalam, and Tamil. The dataset has following two attributes: comments and speech category. Comments denote the opinion of the online users expressed in the form of text and the speech denotes the category of the text, i.e., hope speech or not hope speech. The dataset contained 22,740 comments in English, 4,940 comments in Kannada, 7,873 comments in Malayalam, and 14,199 comments in Tamil. **Table 2** presents distribution of comments in the two categories for each of the four languages. An example instance in English language is as follows:

- "*Injustice is the way the world works. A millionaire paying someone a few dollars to wash his car is injustice*" - Not hope speech
- "*all lives matter, without that we never have peace so to me forever all lives matter*" - Hope speech

Table 2: Number and category distribution of comments for each language.

Language	Hope speech (# of comments)	Not hope speech (# of comments)
English	1962	20778
Tamil	6327	7872
Malayalam	1668	6205
Kannada	1699	3241

For the purpose of carrying out experimental work, each corpus is divided into three sets: train, validation, and test. **Table 3** shows the statistics of the train, test, and validation sets for the hope speech and not hope speech class for comments in each of the four languages.

3.2 Dataset Preparation

The collected dataset undergoes certain pre-processing steps for using it in the experimental work. First, the comments are converted into lower case and the punctuations, special characters (except a-z, A-Z), URLs, stop-words and spaces are removed. The cleaned text is then tokenized and word_index is built from it. These tokens turn into list of sequences. For this the text_to_sequences() method is used. After that padding is done to make the data uniform using pad_sequences() function in Keras.

[3] https://competitions.codalab.org/competitions/

4 Methodology

This section presents details of the methodology adopted for the hope speech detection task. Both, the traditional deep learning-based models and the transformer-based models are implemented. Thereafter, the better performing models are combined in an ensemble configuration.

Table 3: Number of instances in train, validation and test sets for each language.

	English		Tamil		Malayalam		Kannada	
	Hope speech	**Not hope speech**	**Hope speech**	**Not hope speech**	**Hope speech**	**Not hope speech**	**Hope speech**	**Not hope speech**
Train	1589	16830	5125	6376	1351	5025	1376	2625
Valid	177	1870	569	709	150	559	153	292
Test	196	2078	633	787	167	621	170	324

After the data preparation stage (refer Section 3.2), following steps are undertaken to Implement various computational models:

i. *GloVe embedding*: A function is used to read the contents of the GloVe Vector file, which returns a dictionary that maps the words to their respective word embeddings. The maximum length (*maxlen*) of one comment is defined as 512 characters. The embedding matrix assigns zero vector to those words which are not in the GloVe dictionary. Here, the embedding layer is defined using the built-in Keras embedding layer. It maps the words to their embedding vectors from the embedding matrix.

ii. *Deep learning models:* Two deep learning models – Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) are implemented. More details on implementation of these models are provided below. Both the models are trained for 100 epochs with the "Adam" optimizer and "binary cross-entropy" loss function is used for each language.

iii. *Transformer-based models:* Two transformer-based models, namely Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and a variant of that the XLM-RoBERTa are implemented next. Details on implementation of these models is provided below.

iv. *Ensemble model:* An ensemble of better performing individual models is designed after experimenting with deep learning and transformer-based models. It is referred to as Ensemble (LSTM, mBERT, XLM-RoBERTa).

4.1 Deep learning-based models

Two deep neural networks are implemented:
(a) Implementation of Convolutional neural network (CNN)

The convolutional neural network operates over a volume of inputs. Each CNN layer tries to find a pattern or useful information from the data. Here, the input data is processed by convolution layer through extraction of features and application of filters. After applying the filter, an activation function is applied over the intermediate output. To reduce the dimensional complexity, the pooling layer is used in between the convolutional layer. The last layer in CNN is fully connected layer.

Initially, embedding features are passed to the CNN layer consisting of 128 filters. The activation function is rectified linear unit (ReLU). After that MaxPooling layer is used to reduce the dimensional complexity. At the end, dense layer is added with sigmoid function to classify the comments. To mollify the chance of overfitting, dropout technique is used with a dropout rate of 0.2. Dropout is known to be a good regularization technique. **Figure 1** shows the CNN approach to classify the comments either as 'hope speech' or 'not hope speech' for the given data for different languages.

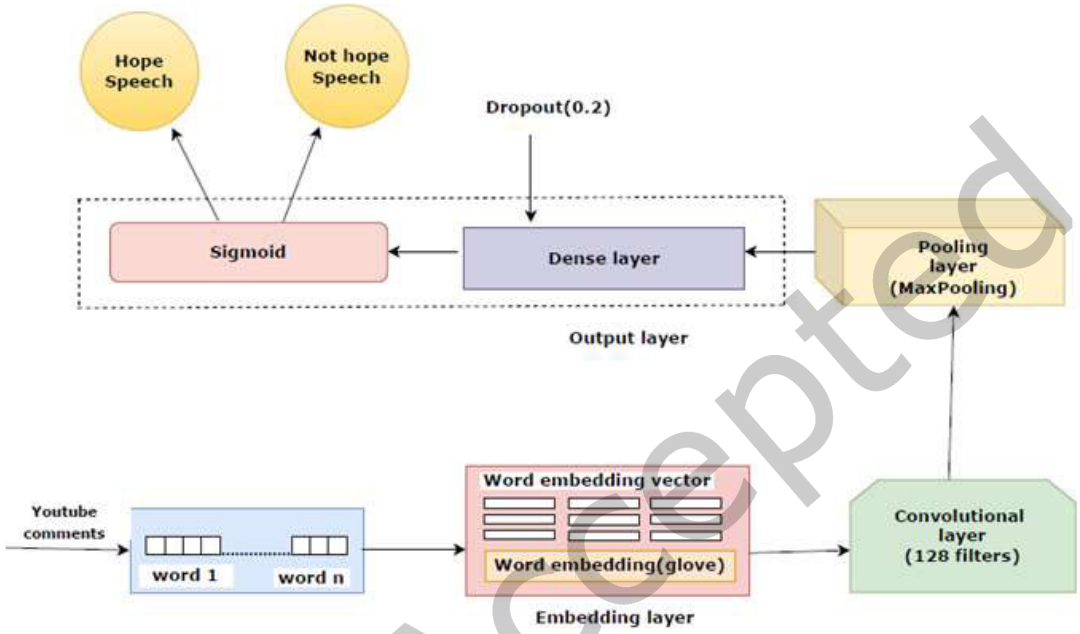


Figure 1: CNN approach to classifying the comments as 'hope speech' or 'not hope speech' for data from different languages.

(b) Implementation of Long Short Term Memory (LSTM)

LSTM is a special kind of Recurrent Neural Network (RNN). It is designed to avoid the long-term dependency problem. LSTM network consists of different memory blocks called cells**.** The cell state and the hidden state are the two states that are being transferred to the next cell. Here, for remembering things and manipulations, the memory blocks are responsible and that is done through three major mechanisms, called gates- Forget gate, Input gate and Output gate.

The obtained embedded vectors are fed to the LSTM model. Embedding matrix assigns zero vector to those words which are not in the GloVe dictionary. Embedding layer is the first layer of the model which uses the max features. LSTM is the next layer with 128 neurons which will work as the memory unit of the model. After that, dense layer is added which is an output layer with sigmoid function, which helps in providing the labels (Hope Speech and Not Hope Speech). **Figure 2** shows the LSTM approach with GloVe embedding to classify comments either as 'hope speech' or 'not hope speech' for the given data for different languages.

4.2 Transformer-based models

Two transformer-based models are implemented for the classification task:

(a) Multilingual Bidirectional Encoder Representations from Transformers (mBERT) Model

BERT is a transformer-based model architecture that learns the contextual relationship between words in the text. A transformer consists of an encoder to read the input text and a decoder gives the prediction for the given task. BERT only needs the encoder part because it aims to generate a language representation model. In this paper,

mBERT base architecture is used that consists of 12 layers of transformer encoder, with each encoder containing sub-layers: a self-attention and a feed-forward layer.

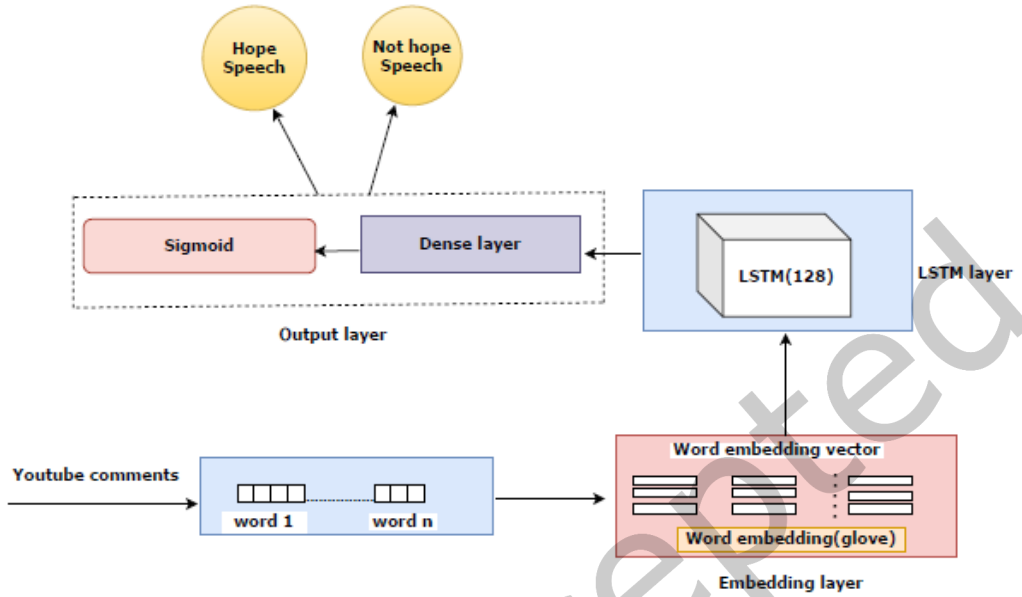


Figure 2: LSTM with GloVe embedding to classify comments into 'hope speech' or 'not hope speech' for data from different languages.

First, the input comments are converted into a sequence of tokens and this process is known as tokenization. A sequence of tokens feeds as an input to the mBERT model. There are two special tokens: [CLS] and [SEP] in each sequence of tokens. The first token of every sequence is [CLS] and [SEP] is used to separate segments. The mBERT tokenizer is used for this purpose. The maximum size of tokens that can be fed into the mBERT model is 512. If the length of tokens in the sequence is less than 512 then the unused token slots are filled by using padding to use [PAD] token and if the length of tokens in the sequence is more than 512 then truncation is needed. **Figure 3** shows how mBERT tokenizer tokenizes the text which is then fed to the mBERT model for classification.

After pre-processing the data, the model is build using the pre-trained BERT-base-multilingual- cased model. Then, the mBERT model outputs embedding vector of size 768 in each of the tokens. The model is trained for 10 epochs with the "AdamW" optimizer and "Binary Cross-entropy" loss function for each language, while the learning rate is set to be 3e-5. **Figure 4** shows the implementation of mBERT to classify the comments for each language.

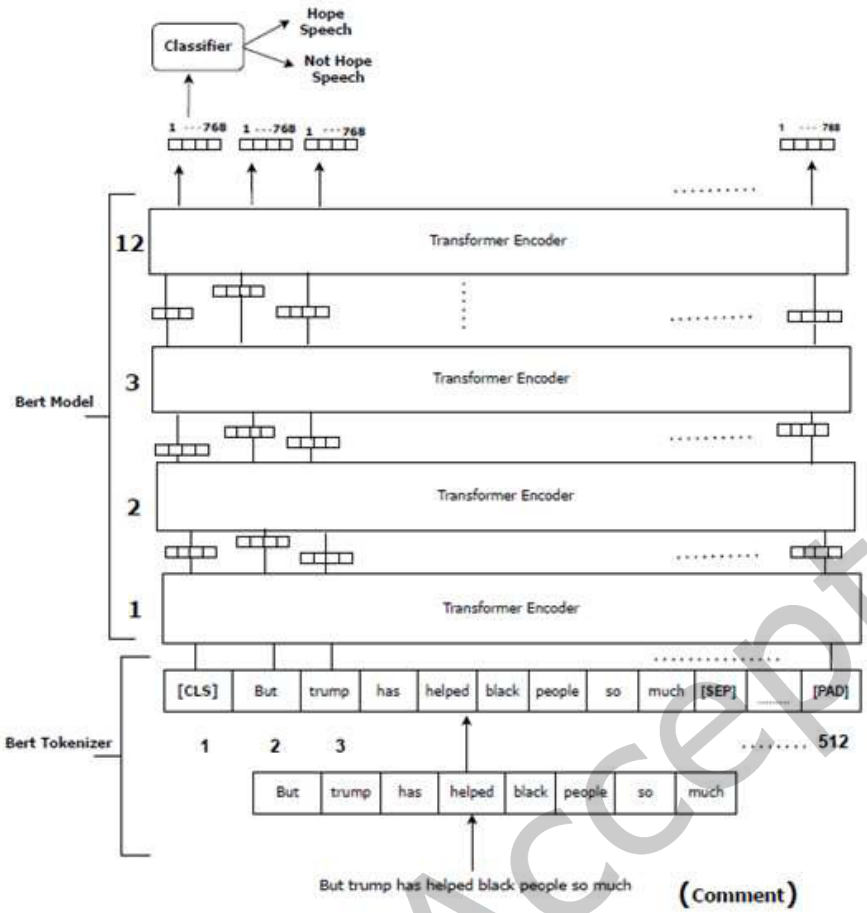


Figure 3: Illustration of mBERT Tokenizer

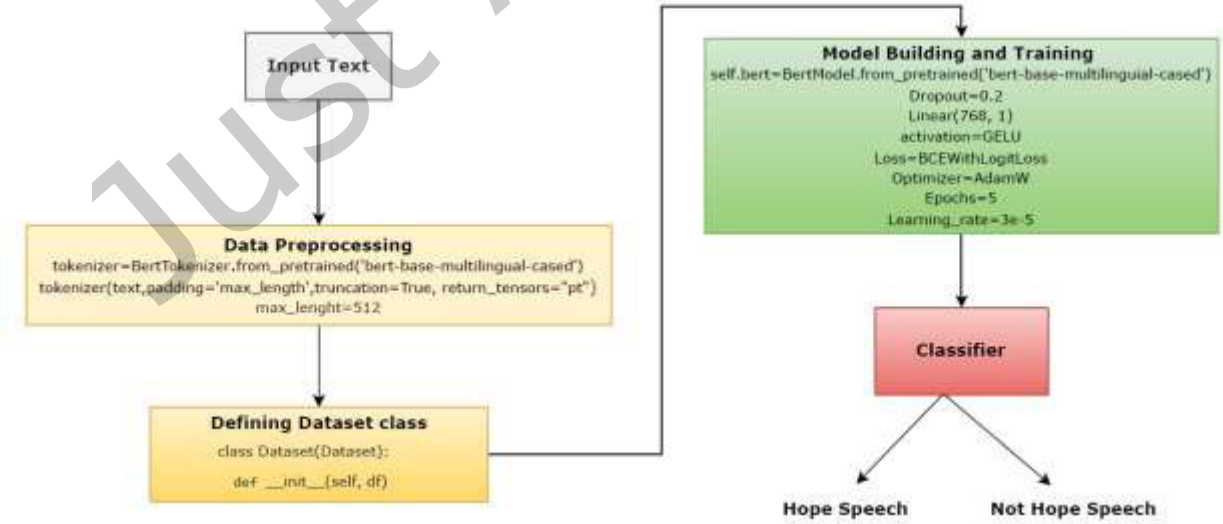


Figure 4: Implementation of mBERT to classify the comments as 'hope speech' or 'not hope speech' for data from different languages.

(b) XLM-RoBERTa

XLM-RoBERTa[4] is based on Facebook RoBERTa model and RoBERTa is based on Google's BERT model [28]. RoBERTa builds on BERT and it modifies key hyperparameters and training with larger mini-batches and learning rates. RoBERTa has the same architecture as BERT. XLM-RoBERTa is pre-trained on 100 languages, so it has powerful vocabulary. Firstly, all punctation like @, 'https://' links, numbers are removed from the comments. XLM-RoBERTa tokenizer is used to tokenize the text. After tokenization, token_ids is generated. The maximum length of the sequence that BERT can processed is 512 with two special tokens: [CLS] and [SEP]. After pre-processing the data, the model is build using the pre-trained xlm-roberta-base model by calling XLMRobertaForSequenceClassification function. Finally, the model is trained for 10 epochs using the "AdamW" optimizer and learning rate set to 3e-5.

4.3 The Proposed Ensemble Model

Ensemble approach is a process in which multiple models are combined and aggregated to predict an output [34], with the idea that when multiple models are combined there is a chance to improve the results. Although different deep learning frameworks and transformer-based models have been implemented for the hope speech detection [24][7][2]. However, a suitable ensemble of appropriate models has not been explored for the task. Based on the performance of individual models, this work proposes an ensemble of deep learning model (LSTM) and two transformer-based models: mBERT and XLM-RoBERTa. The primary reason behind using LSTM model in ensemble approach rather than CNN is that LSTM can entrap the long-term dependencies between word sequences and hence can be better used for text classification [27]. As, the dataset are in Dravidian languages (Malyalam, Kannada, Tamil) along with English, multilingual BERT and XLM-RoBERTa are suitable to classify the comments in English as well as non-English language. Thus, an ensemble model of these three methods is proposed and is denoted by Ensemble (LSTM, mBERT, XLM-RoBERTa). **Figure 5** shows the architectural framework for the model. Ensemble (LSTM, mBERT, XLM-RoBERTa) aggregates the prediction of each individual model involved using a majority voting scheme and renders one final prediction. The model is trained for 10 epochs using the "AdamW" optimizer and the learning rate is set as 3e-5.

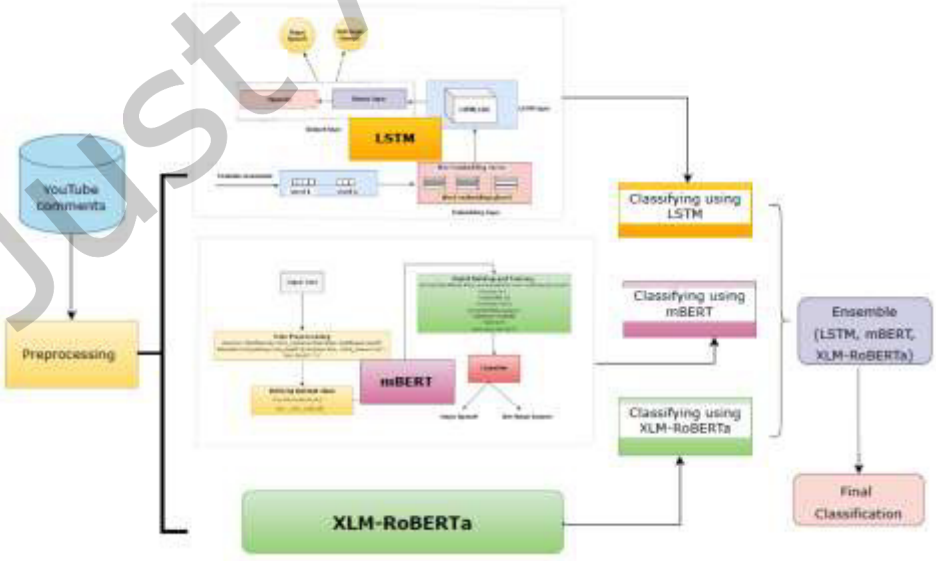


Figure 5: Ensemble model for hope speech detection task.

[4] https://huggingface.co/xlm-roberta-base

5 Results and Comparative Analysis

This section presents results and comparative analysis of the performance of the deep learning, transformer-based and the proposed ensemble models. The performance of the algorithmic models are evaluated using standard metrics and also compared with previous studies.

5.1 Evaluation Metrics

The following standard metric are used to evaluate the performance of the various models implemented.

$$Precision\ (P) = \frac{TP}{(TP\ +\ FP)}$$

$$Recall\ (R) = \frac{TP}{(TP + FN)}$$

$$F1 - score\ = \frac{2 * P * R}{(P + R)}$$

where, TP, FP, and FN are True Positive, False Positive, and False Negative, respectively. The metrics (precision, recall, F1- score) are computed independently for each of the classes and then the weighted average of all is taken and reported. The confusion matrix for data from different languages is considered for computing the metrics.

The **Figure 6** presents the confusion matrix for the proposed ensemble model for the data of the four languages explored. Here, (0,0) indicates TN, (0,1) indicates FN, (1,0) indicates FP and (1,1) indicates TP.

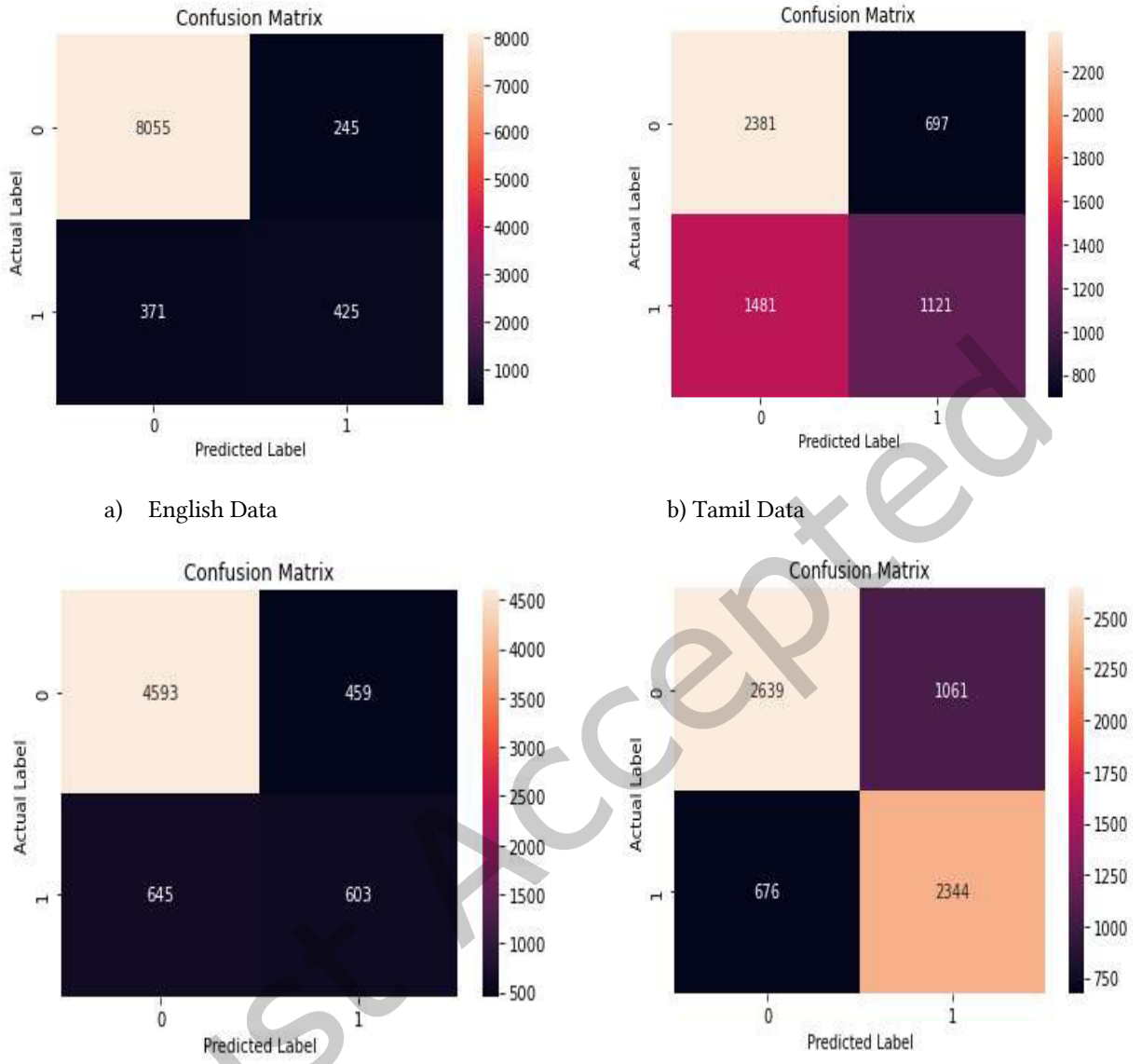
5.2 Experimental Results

Now, the P, R and F1 score values for the different models implemented are presented for data of each of the four languages. **Table 4** shows the performance metrics of the two deep learning models (CNN, LSTM) and transformer-based models (mBERT, XLM-RoBERTa), and the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) for data of the English language. It can be seen that for English language, the weighted average F1-score for CNN, LSTM, mBERT and XLM-RoBERTa are 0.89, 0.90, 0.92 and 0.92, respectively. On the other hand, the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) implementation achieves F1 score of 0.93, which is better than the all the other four implementations. Thus, for English language, the proposed ensemble model outperforms the other models.

Table 4: Classification report for implemented models trained on English data.

Model	**Embedding**	**Hope speech**			**Not hope speech**			**Weighted Average F1-score**
		P	**R**	**F1-score**	**P**	**R**	**F1-score**	
CNN	GloVe	0.43	0.24	0.31	0.93	0.97	0.95	0.89
LSTM	GloVe	0.45	0.28	0.35	0.93	0.97	0.95	0.90
mBERT	-	0.61	0.40	0.48	0.94	0.98	0.96	0.92
XLM-RoBERTa	-	0.66	0.42	0.51	0.95	0.98	0.96	0.92
Ensemble (LSTM,mBERT,XLM-RoBERTa)	-	0.63	0.53	0.58	0.96	0.97	0.96	**0.93**

*P=Precision, R=Recall



a) English Data

b) Tamil Data

c) Malayalam data

d) Kannada data

Figure 6: Confusion matrix of Ensemble (LSTM, mBERT, XLM-RoBERTa) for the four languages considered.

Table 5 shows the performance metrics for the different models implemented on the data of Tamil language. It can be seen that CNN, LSTM, mBERT, XLM-RoBERTa and Ensemble (LSTM, mBERT, XLM-RoBERTa) achieve F1 score of 0.54, 0.57, 0.57, 0.59 and 0.60, respectively. Thus, in case of Tamil language too, the proposed model outperforms the rest of the four models.

Tables 6 and **7** show the performance metrics of the different models implemented on the data of Malayalam and Kannada languages, respectively. From Table 6 it can be seen that Ensemble (LSTM, mBERT, XLM-RoBERTa) outperforms the deep learning models and the other transformer-based models. Similarly, from Table 7, it can be observed that the Ensemble (LSTM, mBERT, XLM-RoBERTa) model achieves better performance. Thus, for data from all the four languages, the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) model achieves better performance than other four models implemented.

Table 5: Classification report for implemented models trained on Tamil data.

Model	Embedding	Hope speech			Not hope speech			Weighted Average F1-score
		P	R	F1-score	P	R	F1-score	
CNN	GloVe	0.49	0.42	0.45	0.58	0.64	0.61	0.54
LSTM	GloVe	0.55	0.41	0.47	0.60	0.73	0.66	0.57
mBERT	-	0.65	0.29	0.41	0.59	0.87	0.70	0.57
XLM-RoBERTa	-	0.68	0.32	0.43	0.60	0.87	0.71	0.59
Ensemble(LSTM,mBERT ,XLM-RoBERTa)	-	0.62	0.43	0.51	0.62	0.77	0.69	**0.60**

*P=Precision, R=Recall

Table 6: Classification report for implemented models trained on Malayalam data.

Model	Embedding	Hope speech			Not hope speech			Weighted Average F1-score
		P	R	F1-score	P	R	F1-score	
CNN	GloVe	0.55	0.23	0.33	0.82	0.95	0.88	0.76
LSTM	GloVe	0.61	0.23	0.33	0.82	0.96	0.89	0.77
mBERT		0.64	0.29	0.40	0.85	0.96	0.90	0.80
XLM-RoBERTa		0.60	0.39	0.47	0.85	0.94	0.90	0.81
Ensemble (LSTM,mBERT,XLM-RoBERTa)		0.57	0.48	0.52	0.88	0.91	0.89	**0.82**

*P=Precision, R=Recall

Table 7: Classification report for implemented models trained on Kannada data.

Model	Embedding	Hope speech			Not hope speech			Weighted Average F1-score
		P	R	F1-score	P	R	F1-score	
CNN	GloVe	0.45	0.61	0.52	0.75	0.60	0.67	0.62
LSTM	GloVe	0.47	0.46	0.47	0.72	0.73	0.73	0.64
mBERT	-	0.65	0.34	0.44	0.72	0.90	0.80	0.67
XLM-RoBERTa	-	0.57	0.51	0.54	0.75	0.80	0.77	0.69
Ensemble(LSTM ,mBERT,XLM-RoBERTa	-	0.69	0.78	0.73	0.80	0.71	0.75	**0.74**

*P=Precision, R=Recall

In order to statistically confirm superior performance of the proposed ensemble model as compared to the other four models, paired sample *t*-test is performed. **Table 8** shows the paired sample *t*-test of the Ensemble (LSTM, mBERT, XLM-RoBERTa) model with other implemented models for hope speech detection. Here, the null hypothesis is that there is no difference between the performance of the proposed ensemble model and other implemented models. In other words, the null hypothesis assumes that the proposed model and other implemented models perform equally. The alternative hypothesis will then be that the proposed model outperforms other applied models in terms of performance. The p-values for the present study (ref. Table 8) are less than the level of significance (α=0.05) in all the cases. This suggests that the null hypothesis can be rejected and hence the alternative hypothesis is true. Thus, statistical test confirms the superior performance of the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) as compared to all the other four models implemented.

Table 8: Paired Sample t-Test of the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) with other implemented models.

Compared Model	t-Test
CNN	t= 2.421 **p= 0.021**
LSTM	t= 2.448 **p= 0.021**
mBERT	t= 2.673 **p= 0.023**
XLM-RoBERTa	t= 2.576 **p= 0.022**

Now that the superior performance of the proposed model is statistically confirmed too, it would be interesting to compare the performance of the proposed model with some previous studies which have used the same dataset for hope speech detection task.

Table 9: Comparison of results of some of the prominent previous studies on the same dataset with the proposed ensemble model

S. No	Authors	Models				Weighted Average F1 score			
		Engl-ish	**Tamil**	**Mala-yalam**	**Kannada**	**Engl-ish**	**Tamil**	**Mala-yalam**	**Kannada**
1.	Dave, B., Bhat, S., & Majumder, P. (2021, April) [4]	TF-IDF (char)+LR	TF-IDF (char)+LR	TF-IDF (char)+LR	-	0.92	0.57	0.72	-
2.	Saumya, S., & Mishra, A. K. (2021, April) [2]	2 parallel CNN-LSTM	3 parallel Bi-LSTM	3 parallel Bi-LSTM	-	0.91	0.56	0.78	-
3.	Arunima, S., Ramakrishnan, A., Balaji, A., & Thenmozhi, D. (2021, April) [6]	BERT	mBERT	mBERT	-	0.92	0.46	0.81	-
4.	Vijayakumar, P., Prathyush, S., Aravind, P., Angel, S., Sivanaiah, R., Rajendram, S. M., & Mirnalinee, T. T. (2022, May) [24]	ALBERT	ALBERT	ALBERT	ALBERT	0.88	0.39	0.74	0.75
5.	Zhu, Y. (2022, May) [13]	Ensemble model (CNN, Bi-LSTM, CNN+LSTM)	Ensemble model (CNN,Bi-LSTM,CNN+LSTM)	Ensemble model (CNN,Bi-LSTM,CNN+LSTM)	Ensemble model (CNN,Bi-LSTM,CNN+LSTM)	0.88	0.41	0.72	0.72
6.	Kumar,A.,Saumya, S., & Roy, P. (2022, May) [23]	Ensemble model (TF-IDF (SVM,LR and Random Forest)	Ensemble model (TF-IDF (SVM,LR and Random Forest)	Ensemble model (TF-IDF (SVM,LR and Random Forest)	Ensemble model (TF-IDF (SVM,LR and Random Forest)	0.88	0.38	0.74	0.75
7.	Proposed Method (Ensemble (LSTM,mBERT,XLM-RoBERTa))	Ensemble(LSTM,mBERT,XLM-RoBERTa	Ensemble (LSTM,mBERT,XLM-RoBERTa	Ensemble (LSTM,mBERT,XLM-RoBERTa	Ensemble(LSTM,mBERT,XLM-RoBERTa	**0.93**	**0.60**	**0.82**	**0.74**

5.3 Comparisons with previous studies

The performance of the proposed model is compared with results of some previous studies which used the same dataset. **Table 9** presents a comparison of the performance of the three major previous studies on the HopeEDI dataset and the proposed Ensemble (LSTM, mBERT, XLM-RoBERTa) model.

Authors in [4] used the machine learning model along with TF-IDF, and obtained average weighted F1 scores of 0.92, 0.57, 0.72 for English, Tamil and Malayalam, respectively. Another study [2] used deep learning models in hybrid manner and proposed models 3-parallel Bi-LSTM and 2-parallel CNN-LSTM. Another study [6] used mBERT for English, Tamil and Malayalam. Authors in [24] used the ALBERT for hope speech detection. Another previous work [13] have proposed the ensemble model consisting of three parts: LSTM, CNN, Bi-LSTM for classifying the comments as 'Hope speech' or 'Not hope speech' in four languages (English, Kannada, Malayalam and Tamil). An ensemble approach that combines a support vector machine, logistic regression, and random forest classifiers is reported in [23]. In the support vector machine and logistic regression classifiers, char-level TF-IDF features were used, whereas in the random forest classifier, word-level features were used. The comparison of results shows that the proposed model- Ensemble (LSTM, mBERT, XLM-RoBERTa) – is superior in performance to other models and previous studies. Further, the obtained results of the proposed model are better than previous studies on the data of all the four languages- English, Kannada, Malayalam and Tamil.

5.4 Error Analysis

The implemented models obtained good accuracy. However, there were some errors of misclassification too. Some such examples are discussed next to understand the probable reasons for misclassification. Examples for data of all the four considered languages (English, Kannada, Malayalam and Tamil) are seen. **Table 10** presents examples of misclassification of data from all the four languages.

Table 10: Some examples of misclassification

Language	Comment and its English translation	Actual Class	Predicted Class
English	i love how people like you read a comment that goes against your agenda and spat instantly zero evidence when all you need to do is type we are trained marxists blm and boom there you go	Non-hope_Speech	Hope_Speech
English	Only one race the Human Race	Hope_Speech	Non-hope_Speech
English	I agree racism is aids. But tearing down the statue is utter bullshit. It is not like people go there to admire him	Hope_Speech	Non-hope_Speech
Tamil	இந்தியா சீனாவை வென்று நம் நாட்டின் பரப்பளவை அதிகரிக்க போகிரது **India is going to conquer China and increase the area of our country**	Hope_Speech	Non-hope_Speech
Tamil	Bro status எடுக்க சிறந்த app எது **Which is the best app to get bro status?**	Non-hope_Speech	Hope_Speech
Tamil	Tik tok ஒளிஞ்சது ஒளிஞதாவே இருக்கட்டும் மது பழக்கத்துக்கு அடிமையானவர்க ளை விட tik tok கு அடிமையானவர்கள் தான் அதிகமாக உள்ளனர் **Let Tik Tok be hidden, there are more Tik Tok addicts than alcoholics.**	Hope_Speech	Non-hope_Speech
Malayalam	ഷർട്ട് itta ചട്ടേന്റെ ചോദ്യം വളരെ നന്നായി കവിന്റെ ഉത്തരം മുട്ടിപോയി **The shirt guy's question was so good that Kevin's answer was knocked over.**	Non-hope_Speech	Hope_Speech
Malayalam	വണ്ട് ശാസ്ത്രം പിന്നോക്കമായിരുന്ന സമയത്ത് മാനവികതയ്ക്കും മനുഷ്യാവകാശത്തിനും നിരക്കാത്ത ഉണ്ടായ തെറ്റുകൾ മുഴുവൻ **All the mistakes made against humanity and human rights**	Hope_Speech	Non-hope_Speech

	when science was backward in the past.		
Malayalam	എന്ത് കണ്ടോളൂ?ഇവരുടെ കാര്യത്തിൽ മാത്രമല്ല.. ഏതൊക്കിലും സെലിബ്രിറ്റികൾ വിവാഹം കഴിക്കുമ് **What do you see? Not only in their case.. any celebrities get married**	Hope_Speech	Non-hope_Speech
Kannada	By mistakly nimma ondu video nodi subscribe agbitte keep it up bro. **By mistakely watch one of your videos subscribe and keep it up bro**	Non-hope_Speech	Hope_Speech
Kannada	Binduge saryagi ugithidira good go ahead we are enjoying well **Bindu saragi ugitidira good go ahead we are enjoying well.**	Non-hope_Speech	Hope_Speech
Kannada	ನೀವು ಹೇಳಿದ ಕಥೆ ಚನನಾಗಿಯೇ ಇದೆ ಆದರೆ ಅಶವತತಾಮ ಬದುಕೆರೋದು ಸುಳ್ಳು **The story you told is good but Ashwatthama's life is a lie**	Hope_Speech	Non-hope_Speech

It can be seen that most of the wrong predictions are because of the presence of certain words in the comment that are used beyond their literal meaning. For example, consider the following comment from the English dataset:

Text: "i love how people like you read a comment that goes against your agenda and spat instantly zero evidence when all you need to do is type we are trained marxists blm and boom there you go".

In this comment, some words such as 'love', 'like' and 'boom' which are usually used for denoting support, motivation, or encouragement, resulting in a misclassification. Comments which have sarcasm or have some implicit/ hidden meanings are difficult to classify.

For Tamil language, it can see that many comments are misclassified for various reasons that the model cannot detect. Consider for example the following comment:

Text: இந்தியா சீனாவை வென்று நம் நாட்டின் பரப்பளவை அதிகரிக்க போகிரது

Translation: India is going to conquer China and increase the area of our country

This comment does not contain representative terms for hope speech. There are other similar examples of misclassification.

The performance of all the models on the Malayalam dataset is commendable, as most of the models achieved weighted average F1-Scores greater than 0.75. The best-performing model misclassified relatively lesser examples in contrast to its performance on other Dravidian languages (Tamil, Kannada) datasets. One such example is:

Text: ഷർട്ട് itta ചട്ടേന്റെ ചോദ്യം വളരെ നന്നായി കവിന്റെ ഉത്തരം മുട്ടിപോയി

Translation: The shirt guy's question was so good that Kevin's answer was knocked over.

The label of this comment is 'not_hope_speech' but the model predicted it as hope speech. The reason for misclassification may be because of positive words used for appreciation. Therefore, the model incorrectly predicted the comment as 'hope_speech'.

A similar example can be taken for the data from Kannada language as well:

Text: By mistakly nimma ondu video nodi subscribe agbitte keep it up bro.

Translation: By mistakely watch one of your videos subscribe and keep it up bro.

This comment sounds like a 'hope speech'' but the appreciation is sarcastic, and hence the the model could not identify the sarcasm resulting in misclassification.

6 Discussion

The results provide important insights into the detection of hope speech across four languages: English, Kannada, Malayalam, and Tamil. A combination of deep learning (LSTM) and transformer-based models (mBERT, XLM-RoBERTa) was used and it was found that the ensemble model consistently outperformed individual models, particularly when dealing with low-resource languages.

The ensemble model performed best for English (F1 score: 0.93), demonstrating the strength of transformer-based architectures when ample data is available. However, for low-resource languages like Kannada (F1 score: 0.74), Malayalam (F1 score: 0.82), and Tamil (F1 score: 0.60), the model encountered more challenges. It struggled with issues like informal language, sarcasm, and cultural differences (discussed in Section 5.4). The lower performance for Tamil is likely due to its complex sentence structure and the limited amount of labeled data available. This suggests that better results for Tamil would require more specialized models or additional training data.

The fact that the ensemble model outperformed standalone models (CNN, LSTM, mBERT, XLM-RoBERTa) across all languages highlights the benefits of combining different model architectures. LSTM, for example, excels at capturing long-term dependencies in text, which is essential for understanding hope speech in longer comments. On the other hand, transformer models like mBERT and XLM-RoBERTa are particularly strong at understanding the context of individual words in multilingual texts. These results suggest that while transformer models are powerful, combining them with deep learning models like LSTM, which specialize in handling sequences, offers a more comprehensive understanding of the text, leading to better classification accuracy. **Table 11** shows the training hyperparameters of proposed ensemble model were kept the same for all languages ensuring a fair comparison.

Table 11: Details of Hyperparameters used to train Ensemble model

Hyperparameters	**Value**
Train Test Validation Split	80:10:10
Batch size	32
Loss Function	Binary Cross-entropy
Optimizer	AdamW
Learning rate	3e-5
Activation Function	Sigmoid
Epochs	10
Evaluation Metric	Weighted Average F1-score

This study also underscores the difficulties of working with low-resource languages, where high-quality labeled data is scarce. For languages like Kannada, Malayalam, and Tamil, this limitation resulted in lower performance compared to English. Additionally, the lack of linguistic tools such as tokenizers and word embeddings for these languages further complicates the task. The findings suggest that regional language models or the use of multilingual pre-trained models can improve performance in future research. A key limitation across

all languages was the model's difficulty in detecting sarcastic or subtly negative comments that appeared positive on the surface. This led to some misclassifications, as seen in the error analysis (section 5.4), where phrases containing positive words were labeled as hope speech, despite their sarcastic intent. This highlights a broader challenge: models need to become better at detecting the nuances of language, such as sarcasm and figurative speech.

7 Conclusion

The study presents experimental work towards design of suitable algorithmic models for hope speech detection from You Tube comments in four different languages- English, Kannada, Malayalam and Tamil. Different standalone models are explored, and a new ensemble model is proposed for the task (RQ1). Results obtained show that the proposed ensemble model has outperformed the individual models (CNN, LSTM, mBERT, XLM-RoBERTa) as well as models proposed in earlier studies (RQ2). This may be attributed to the fact that the constituent models are trained on different corpus and therefore combining the models provides for a notably large and heterogeneous training. In this way, the ensemble model can combine the knowledge of individual models together for achieving better performance. Further, combining the models allows for fine-tuning and adjustment of various hyperparameters (learning rate, epsilon and number of epochs), which also positively impacts the model performance. The proposed model is found to perform well on the data from all the four languages. Therefore, the research work adds to the knowledge in the area in the form of a new model which outperforms the existing models.

There are, however, certain limitations of the present work which also indicate towards some future work possibilities. *First*, the model has been tested on four languages only and its applicability on other languages can be demonstrated subject to availability of suitable datasets in those languages. *Second*, the techniques of balancing the dataset to further improve the performance of the proposed model are not applied in the current work and can be incorporated in future work. *Third*, domain and language specific inputs are not considered in the present case and hence one may try incorporating the pragmatics of the language and any domain specific knowledge available to further improve the model. *Fourth*, the proposed model is not able to correctly deal with situations of implicit/ hidden meanings and sarcasm in the comments. Therefore, more studies are required to explore these aspects. *Finally*, the present work did not explore on the explainability of the proposed, which can be taken up as a future work. Overall, the study presents a useful contribution in form of experimental work and a proposed model for hope speech detection from texts in different languages. It is important to have such an algorithmic model to automatically identify what comments constitute 'hope speech' so that suitable strategies can be devised to promote such content on social media that are supportive, enjoyable and can motivate people in a positive way, especially those suffering from depression, distress etc. More research work in this direction can be useful for social media content monitoring and analysis.

ACKNOWLEDGMENTS

This work is partly supported by HPE Aruba Centre for Research in Information Systems, Banaras Hindu University, Varanasi, India (Project Code- M-22-69) to the third author. 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).

REFERENCES

[1] Singh, A., Sharma, D., & Singh, V. K. (2024). MIMIC: Misogyny Identification in Multimodal Internet Content in Hindi-English Code-Mixed Language. ACM Transactions on Asian and Low-Resource Language Information Processing.

[2] Saumya, S., & Mishra, A. K. (2021, April). IIIT_DWD@ LT-EDI-EACL2021: hope speech detection in YouTube multilingual comments. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 107-113).

[3] Jha, V., Mishra, A., & Saumya, S. (2022, May). CURAJ_IIITDWD@ LT-EDI-ACL 2022: Hope Speech Detection in English YouTube Comments using Deep Learning Techniques. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 190-195).

[4] Dave, B., Bhat, S., & Majumder, P. (2021, April). IRNLP_DAIICT@ LT-EDI-EACL2021: hope speech detection in code mixed text using TF-IDF char n-grams and MuRIL. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 114-117).

[5] Chakravarthi, B. R. (2020, December). 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* (pp. 41-53)

[6] Arunima, S., Ramakrishnan, A., Balaji, A., & Thenmozhi, D. (2021, April). ssn_diBERTsity@ LT-EDI-EACL2021: hope speech detection on multilingual YouTube comments via transformer based approach. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 92-97).

[7] Singh, P., Kumar, P., & Bhattacharyya, P. (2021, April). CFILT IIT Bombay@ LT-EDI-EACL2021: Hope Speech Detection for Equality, Diversity, and Inclusion using Multilingual Representation fromTransformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 193-196).

[8] Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security* (pp. 2-12).

[9] Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., & Chakravarthi, B. R. (2021). IIITT@ LT-EDI-EACL2021-hope speech detection: there is always hope in transformers. *arXiv preprint arXiv:2104.09066*.

[10] Chen, S., & Kong, B. (2021, April). cs_english@ LT-EDI-EACL2021: Hope Speech Detection Based On Fine-tuning ALBERT Model. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 128-131).

[11] Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., Cn, S., McCrae, J. P., García, M. Á., Jiménez-Zafra, S.M., Valencia-García, R., Kumaresan, P., Ponnusamy, R., & García-Díaz, J. (2022, May). Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 378-388).

[12] Chakravarthi, B. R., & Muralidaran, V. (2021, April). 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* (pp. 61-72).

[13] Zhu, Y. (2022, May). Lps@ lt-edi-acl2022: an ensemble approach about Hope Speech Detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 183-189).

[14] Schmidt, A., & Wiegand, M. (2019, January). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain* (pp. 1-10). Association for Computational Linguistics.

[15] Mondal, M., Silva, L. A., & Benevenuto, F. (2017, July). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media* (pp. 85-94).

[16] Balouchzahi, F., Aparna, B. K., & Shashirekha, H. L. (2021, April). MUCS@ LT-EDI-EACL2021: coHope-hope speech detection for equality, diversity, and inclusion in code-mixed texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 180-187).

[17] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

[18] Singh, A., Sharma, D., & Singh, V. K. (2025). Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models. Computer Speech & Language, 89, 101682.

[19] Warner, W., & Hirschberg, J. (2012, June). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19-26).

[20] Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, *10*(12), 4180.

[21] Que, Q. (2021, April). Simon@ LT-EDI-EACL2021: Detecting Hope Speech with BERT. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 175-179).

[22] Hande, A., Hegde, S. U., Sangeetha, S., Priyadharshini, R., & Chakravarthi, B. R. (2022, May). The Best of both Worlds: Dual Channel Language modeling for Hope Speech Detection in low-resourced Kannada. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 127-135).

[23] Kumar, A., Saumya, S., & Roy, P. (2022, May). SOA_NLP@ LT-EDI-ACL2022: An Ensemble Model for Hope Speech Detection from YouTube Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 223-228).

[24] Vijayakumar, P., Prathyush, S., Aravind, P., Angel, S., Sivanaiah, R., Rajendram, S. M., & Mirnalinee, T. T. (2022, May). SSN_ARMM@ LT-EDI-ACL2022: Hope Speech Detection for Equality, Diversity, and Inclusion Using ALBERT model. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 172-176).

[25] Sharma, D., Singh, V. K., & Gupta, V. (2024). TABHATE: a target-based hate speech detection dataset in Hindi. Social Network Analysis and Mining, 14(1), 190.

[26] Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021, April). Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 149-159).

[27] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

[28] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[29] Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application*, 267-281.

[30] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 512-515).

[31] Balouchzahi, F., Butt, S., Sidorov, G., & Gelbukh, A. (2022, May). CIC@ LT-EDI-ACL2022: Are transformers the only hope? Hope speech detection for Spanish and English comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 206-211).

[32] Gupta, V., Kumar, R., & Pamula, R. (2022, May). IIT Dhanbad@ LT-EDI-ACL2022-Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 229-233).

[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[34] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

[35] Sharma, D., Gupta, V., & Singh, V. K. (2022, December). Detection of homophobia & transphobia in Malayalam and Tamil: Exploring deep learning methods. In *International Conference on Advanced Network Technologies and Intelligent Computing* (pp. 217-226). Cham: Springer Nature Switzerland.

[36] Bansal, V., Tyagi, M., Sharma, R., Gupta, V., & Xin, Q. (2022). A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[37] Sharma, D., Singh, A., & Singh, V. K. (2024). THAR-Targeted Hate Speech Against Religion: A high-quality Hindi-English code-mixed Dataset with the Application of Deep Learning Models for Automatic Detection. ACM Transactions on Asian and Low-Resource Language Information Processing.

[38] Gupta, V., Jain, N., Shubham, S., Madan, A., Chaudhary, A., & Xin, Q. (2021). Toward integrated CNN-based sentiment analysis of tweets for scarce-resource language—Hindi. *Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1-23.

[39] Chakravarthi, B. R. (2022). Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, *12*(1), 75.

[40] García-Baena, D., García-Cumbreras, M. Á., Jiménez-Zafra, S. M., García-Díaz, J. A., & Valencia-García, R. (2023). Hope speech detection in Spanish: The LGBT case. *Language Resources and Evaluation*, 1-28.