Sentiment Analysis of Code-Mixed Tamil and Tulu by Training Contextualized ELMo Representations

Toqeer Ehsan[1], Amina Tehseen[2], Kengatharaiyer Sarveswaran[3], Amjad Ali[4]
[1]Department of Computer Science, University of Gujrat, Pakistan
[2]Department of Information Technology, University of Gujrat, Pakistan
[3]Department of Computer Science, University of Jaffna, Sri Lanka
[4]Information and Computing Technology (ICT) Division, College of Science and Engineering (CSE),
Hamad Bin Khalifa University, Doha, Qatar
toqeer.ehsan@uog.edu.pk, amina.tehseen@outlook.com,
sarves@univ.jfn.ac.lk, amsali@hbku.edu.qa

Abstract

Sentiment analysis in natural language processing (NLP), endeavors to computationally identify and extract subjective information from textual data. For low-resourced languages such as Tamil and Tulu, predicting sentiment becomes a challenging task due to the presence of text comprising various scripts and langauges. In this research, we present the sentiment analysis of code-mixed Tamil and Tulu YouTube comments. We have developed Bidirectional Long-Short Term Memory (BiLSTM) networks based models for both languages by deploying contextualized word embeddings at input layers of the models. For that purpose, ELMo embeddings have been trained on larger unannotated code-mixed text like corpora. Our models performed with macro average F_1-scores of 0.2877 and 0.5133 on Tamil and Tulu code-mixed datasets respectively.

1 Introduction

Sentiment analysis, a subfield of Natural Language Processing (NLP), pursuits to computationally identify and extract subjective data, such as opinions, emotions, and attitudes from textual data. It plays a crucial role in understanding human's evaluations expressed in numerous forms of communication, including social media, customer reviews, and online forums (Thavareesan and Mahesan, 2020). The proliferation of social media platforms allows individuals to proportion their perspective public opinions in written form on the internet (Patra et al., 2018). The users having knowledge of multiple languages, often post their thoughts and reaction in multilingualism. It happens due to no restrictions or limitations on the usage of diverse languages or their syntactic rules (Suryawanshi et al., 2020).

The practice of blending multiple languages at various levels, including sentences, words, or subwords within the same text, is known as code-mixing. There are several reasons for code-mixing such as bilingualism, social community, vocabulary, the speaker and their conversation partner, the context or situation, and social prestige (Balahur and Turchi, 2014). These are considered the primary factors influencing code-mixing on social media networks. Code-mixing often occurs due to the unavailability of a particular word or phrase in a particular language, compelling individuals to incorporate words or phrases from their native language in order to enhance comprehension for the receiver (Ahmad et al., 2022).

Although sentiment analysis has gained significant attention in recent years, most of the research has primarily focused on monolingual text, predominantly in English. However, the emergence of code-mixed text brings forth distinctive challenges and opportunities for researchers. The developing presence of code-mixed text presents unique demanding situations and possibilities for sentiment analysis. There is a limited amount of research on sentiment analysis in low-resourced languages particularly for Tamil and Tulu. The datasets for this research, contains various types of languages including Tamil, Tulu, English, Romanized Tamil, and Tulu, as well as mixed text and emoticons. The diverse range of languages in the text presents significant difficulties in achieving higher accuracy of sentiment prediction models.

In this paper, we present the sentiment analysis of code-mixed Tamil and Tulu YouTube comments as a shared task[1]. We propose Bidirectional Long-Short Term Memory (BiLSTM) networks based models for both languages; Tamil and Tulu which further use contextualized word embeddings at input layers of the models. For that purpose, Embeddings from Language Models (ELMo) em-

[1] https://codalab.lisn.upsaclay.fr/competitions/11095

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 152–159, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_020

beddings have been trained on larger unannotated code-mixed like corpora. For both language, the transfer learning by using trained ELMo models was quite helpful to achieve the improved sentiment prediction results. Our models performed with the macro average F1-scores of 0.2877 and 0.5133 on Tamil and Tulu code-mixed datasets respectively. ELMo Embeddings have shown state of the art performances for low-resourced NLP tasks such as part of speech tagging (Tehseen et al., 2022), phrase chunking (Ehsan et al., 2022), constituency (Ehsan and Hussain, 2020, 2022) and dependency parsing (Ehsan and Butt, 2020). The next sections of the paper present literature review, corpora details, model architecture and results.

2 Literature review

Tamil, being one of the ancient languages with a vibrant literary heritage, is predominantly spoken in the Indian state of Tamil Nadu and certain regions of Sri Lanka (Chakravarthi et al., 2018). There has been a surge in interest regarding sentiment analysis in Tamil, given its extensive usage in diverse domains such as social media, news, and product reviews. Numerous research studies have concentrated on constructing sentiment analysis models specifically for Tamil, employing a range of methodologies, including rule-based techniques, machine learning algorithms, and deep learning architectures.

For code-mixed Tamil-English text's sentiment analysis, Chakravarthi et al. (2020a) developed a corpus, TamilMixSentiment [2], which is a corpus comprising Tanglish (a mix of Tamil and English) comments from YouTube videos. The development of TamilMixSentiment followed guidelines based on the work from Mohammad (2016) and without annotating language tags at the word-level. The inter-annotator agreement was found to be 0.6, indicating a moderate level of agreement among the annotators. They annotated 15,744 comments, making it the largest sentiment corpus for the under-resourced language featuring code-mixing phenomena. They detailed the procedure of developing a code-mixed corpus and attributing polarities. Further, they presented the outcomes of sentiment analysis trained on the corpus, serving as a benchmark.

Chakravarthi et al. (2020b) employed BiLSTM and Recurrent Neural Networks (RNN) with subword representation to categorize text based on its polarity. Additionally, for code-mixed Tamil-English corpus, Chakravarthi et al. (2022) introduced three sentiment assessment frameworks: BERT (Bidirectional Encoder Representations from Transformers) and logistic regression classifier, DistilBERT, and rapid Text-mod.

For Tamil code-mixed sentiment analysis, Shanmugavadivel et al. (2022) analyzed machine learning frameworks. The research objective was to develop hybrid deep learning models that combine Convolutional Neural Network (CNN) with LSTM and CNN+BiLSTM. Hybrid models performance was compared with state-of-the-art methods, including traditional machine learning techniques. Among all their developed models, the proposed CNN+BiLSTM framework outperformed with an accuracy of 66%.

Tulu is an Indian language belonging to the Dravidian language family, spoken mainly in the region. It is gaining attention in sentiment analysis research. However, compared to other languages, Tulu has not been extensively studied in this area. The limited focus on sentiment analysis in Tulu can be attributed to the lack of annotated datasets and linguistic resources available for the language.

For sentiment analysis of Tulu-English code-mixed text, Kannadaguli (2021) developed a corpus comprising 5,536 YouTube comments. The dataset construction focused on extracting comments written in the Latin script of Tulu and Tulu-English code-mixed. The annotated Tulu-English dataset was then utilized to implement various machine learning (ML) and deep learning models, and a transformer-based classifier using BERT framework. Keras embeddings and Term-Frequency-Inverse Document-Frequencies (TF-IDF) were used as attributes for deep learning and machine learning models respectively. The BiLSTM framework demonstrated the best performance with notable F_1-scores across all the classes.

Hegde et al. (2022a) worked on corpus creation for code-mixed Tulu Text for sentiment analysis. They scraped 7,171 YouTube comments and subsequently annotated them to predict emotions within the code-mixed Tulu data, establishing a foundational benchmark. They utilized traditional ML algorithms employing TF-IDF features derived from word bigrams and trigrams. In all sentiment classes, the Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers performed comparably better and reported an F_1-score of 0.60.

[2] https://github.com/bharathichezhiyan/TamilMixSentiment

3 Code-Mixed Corpora

In this section, we present details of corpora and datasets which have been used to train the sentiment analysis and transfer learning. The train, development and test sets were released by the organizers of the shared task. Moreover, we used additional corpora to train ELMo embeddings to perform transfer learning for both Tamil and Tulu code-mixed text.

3.1 Tamil

Details of the Tamil code-mixed text for training and evaluation sets are given in the Table 1. The number of sentences in the Tamil dataset is greater than that in the Tulu dataset. Tamil train set contains 320,746 tokens in total with 9.4 tokens per sentence on average.

Category	**# Sentences**	**# Tokens**
Training set	33,989	320,764
Development set	3,786	35,424
Test set	649	8,019

Table 1: Details of Tamil code-mixed train, development and test sets.

The sentences which are actually YouTube comments, also have emoticons of different types. As the text contains sentences from different scripts and hence has large vocabulary. The Romanized text usually lacks the standard spellings which makes it challenging to train and achieve better predictions. The dataset contains sentences in Tamil script, Romanized Tamil, English, Tamil and Romanized and Tamil with English phrase. Each sentence has been categorized in any of the four sentiment classes; Positive, Negative, Mixed Feelings and Unknown State.

The transfer learning is a suitable method for small to medium sized annotated datasets. We have trained contextualized ELMo embeddings to achieve context-sensitive word vectors and to cater Out-Of-Vocabulary (OOV) words. The ELMo embeddings produce character-based word vectors which are helpful to learn morphology as well as semantics of a language. To perform the transfer learning for code-mixed Tamil, we have used corpora from multiple sources containing text in various scripts. The Table 3 shows statistics of corpora from different sources. The Tamil text corpus has been collected from Kaggle[3], the repository of Tamil - Language Corpus for NLP. We have used a sub-directory containing 357 files of Tamil text with 1,444,046 sentences and 50,743,745 tokens. The Romanized Tamil corpus has been collected from CC100 corpora[4] (Conneau et al., 2019). The corpus contains 6,243,679 sentences and 36,893,050 tokens. The English language text has been collected from Kaggle under the repository IMDB 320.000 Movie Reviews[5]. The repository contains reviews from IMDB[6]. The IMDB dataset contains 320,748 comments and 83,249,225 tokens. As the training dataset has been collected from YouTube comments, therefore, the movie reviews corpus will be useful to perform transfer learning. An additional News and Blogs[7] corpus has been collected from Kaggle to increase the size of the English corpus. The News and Blogs dataset contains 160,036 sentences and 82,627,416 tokens. Overall, the corpus contains 8,168,509 sentences and 253,513,436 tokens from the four sources.

3.2 Tulu

Details of the Tulu code-mixed text provided for training and evaluations are given in the Table 2. Tulu is a low-resourced language with a scarce online corpora. In the shared task, Tulu language has only 6,457 sentences and 36,628 tokens. Similarly, the development and test sets contain 4,729 and 4,077 tokens respectively. The dataset contains sentences in Tulu script, Romanized Tulu, Tulu with Romanized and English text. There are four sentiment classes, which are; Positive, Negative, Neutral and Mixed Feelings.

Category	**# Sentences**	**# Tokens**
Training set	6,457	36,628
Development set	781	4,729
Test set	708	4,077

Table 2: Details of Tulu code-mixed train, development and test sets.

To overcome the data scarcity of Tulu text resources, we have performed the transfer learning by using Kannada language corpus. As Tulu is closely related to Kannada with respect to vocabulary and linguistic features (Hegde et al., 2022b; Vyawahare

[3] https://www.kaggle.com/datasets/praveengovi/tamil-language-corpus-for-nlp

[4] https://metatext.io/datasets/cc100-tamil-romanized

[5] https://www.kaggle.com/datasets/nikosfragkis/imdb-320000-movie-reviews-sentiment-analysis

[6] https://www.imdb.com/

[7] https://www.kaggle.com/datasets/patjob/articlescrape

Sr.#	Data source	# Sentences	# Tokens
1	Tamil - Language Corpus for NLP	1,444,046	50,743,745
2	CC100 - Tamil Romanized[8]	6,243,679	36,893,050
3	IMDB 320.000 Movie Reviews	320,748	83,249,225
4	News and Blog	160,036	82,627,416
	Total	8,168,509	253,513,436

Table 3: Details of Tamil code-mixed corpora to perfrom transfer learning by training contextualized ELMo embeddings.

Sr.#	Data source	# Sentences	# Tokens
1	CC100 - Kannada	2,000,000	30,376,315
2	CC100 - Kannada (Romanized)	2,000,000	30,376,315
3	IMDB 320.000 Movie Reviews	320,748	83,249,225
	Total	4,320,748	144,001,855

Table 4: Details of Tulu code-mixed corpora to perform transfer learning by training contextualized ELMo embeddings.

et al., 2022). We further transliterated the Kannada text to Romanized Kannada by using a transliterator called *om-transliterator*[9]. The transliterator is an open source python library which is freely available to use. Table 4 shows statistics of corpora which have been used to train ELMo embeddings. The Kannada corpus has been collected from CC100 corpora[10] (Conneau et al., 2019). We have used two million sentences from this corpus which were further transliterated to the Roman script. Additionally, the IMDB 320.000 Movie Reviews dataset has been included to the code-mixed large corpus for transfer learning.

3.3 Corpus Preparation

A few pre-processing operations have been performed on the labeled datasets as well as unlabeled corpora. In both Tamil and Tulu datasets, Romanized and English text has been converted to lower case. Tokenization is a basic task to perform any NLP task. Initially, the tokenization has been performed on the basis of the space character but there are many non-word tokens like punctuation and symbols which were combined with words. We have separated these types of symbols from words and used them as separate tokens. Both Tamil and Tulu datasets contain emoticons in them which are quite important to represent the sentiments and feelings. In many sentences, there are repeating emoticons without space in them. We have converted all emoticons to the English text in both datasets. For that purpose, we have used a python package *emoji*[11] which has a function to *demojize* the text. This function takes a sentence as input and returns the same sentence by replacing emoticons with equivalent text. The *demojization* was quite helpful to learning contextual word vectors to learn the sentiment labels.

4 Model

We developed the sentiment analysis model by using Bidirectional Long-Short Term Memory (BiLSTM) networks. BiLSTM based neural models are quite capable to learn sequence labels which can also be used to predict sentence level tasks like sentiment analysis. The model has two LSTM layers, first layer scans the word sequences in the forward direction while the other layer scans the word sequences in opposite i.e. backward direction. The LSTM based models learn next and previous words to attain the contextual information within sentences. The input sequence of N words $x_1,x_2,...,x_n$ is given as input. The BiLSTM($\mathrm{x}_{1:n}$,i) function has been shown in the Equation 1 which demonstrates the concatenation of forward and backward layers. LSTM$_\mathrm{f}$ represents the forward layer whereas LSTM$_\mathrm{r}$ shows the backward layer. The function denotes to a vector *i* by conditioning the past antiquity $\mathrm{x}_{1:i}$ and the forthcoming sequence $\mathrm{x}_{i:n}$ as well.

$$BiLSTM(x_{1:n}, i) = LSTM_f(x_{1:i}).LSTM_r(x_{n:i}) \quad (1)$$

[9] https://pypi.org/project/om-transliterator

[10] https://metatext.io/datasets/cc100-kannada

[11] https://pypi.org/project/emoji

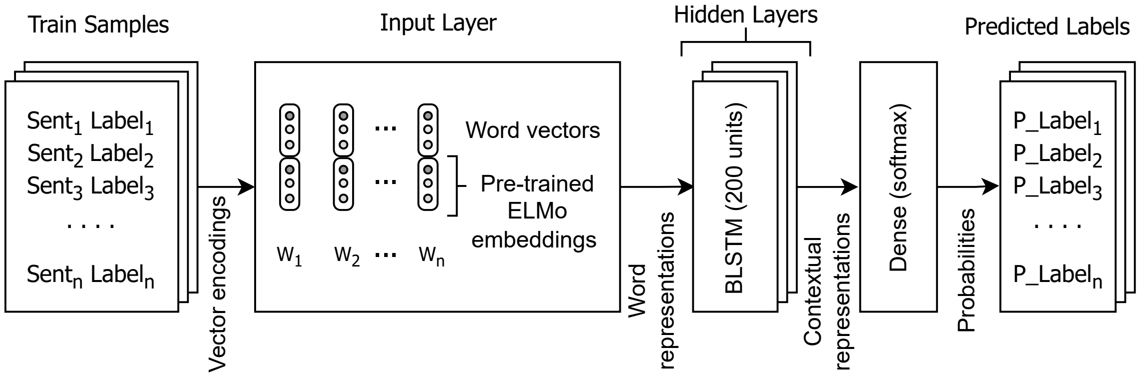

Figure 1: BiLSTM based sentiment analysis model architecture.

The *softmax* non-linearity function is employed at the output layer which performs the multi-class classification returning sentiment label classes l_i for each input sequence of N words $x_1,x_2,...,x_n$ as shown in Equation 2.

$$o_i = Softmax(Xh_i + b) \tag{2}$$

The Figure 1 shows our BiLSTM based neural model architecture which has been trained for sentiment analysis for the Tamil and Tulu code-mixed datasets. The dataset contains sentences followed by sentiment labels. The training sentences have been transformed to vectors to use them in the neural model at input layer. The word vectors have been concatenated with the contextualized vectors achieved from the EMLMo embeddings. The concatenated vectors are further fed to hidden LSTM layers. We experimented with two hidden LSTM layers. The *Dropout* layers have been added between LSTM layers and before the output layer. The LSTM layers are followed by a *Flatten* layer to perform sentiment labeling. The *Softmax* non-linearity has been used at dense output layer to perform multi-class classification. Finally, the sentiment classes are predicted on the basis of maximum likelihood. The predicted labels have been evaluated against gold labels for development and test sets.

The sentiment analysis model has been developed by using *keras* library with the *Tensorflow* back-end in Python-3. The bidirectional LSTM layers had 200 hidden units. The value for *Dropout* layers was set to 0.2 (20%). Root Mean Squared Propagation (*RMSprop(0.001)*) optimization function has been used in the model. The *categorical cross-entropy* loss function was used in all experiments. The word vectors have been trained to have 128 dimensions, however, ELMo embeddings have been trained with 256 projection dimensions. The LSTM sentence length was set to the longest sentence with padding sequences in the datasets for both Tamil and Tulu. Transfer leaning by training ELMo embeddings were quite useful to achieve better results. The following section describes the details of ELMo embeddings and its parameters.

4.1 Transfer Learning

Deep learning based models are data hungry models as they require a lot of annotated samples to produce the state-of-the-art results. However, the annotation of such huge datasets is quite costly in terms of human resource, expertise and time. Transfer learning is a suitable technique by training word representations on large unannotated corpora. This method helps in the training by learning context and OOV tokens. We have trained ELMo embeddings on code-mixed text for Tamil and Tulu languages.

Context-free word embeddings produce unique vectors for each token in the corpus which represent a single meaning. The well-known context-free word embeddings are GloVe (Pennington et al., 2014), Word2Vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017). However, contextualized word representations are able to learn meanings with respect to the contexts of the words because a single word may have multiple meanings in a language. ELMo embeddings (Peters et al., 2018) produce contextualized vectors to learn the meanings with respect to the context. The code-mixed datasets contain Romanized text for both Tamil and Tulu. People usually use the Roman spellings according to their personal practices which results in a lot of variations in the text producing larger

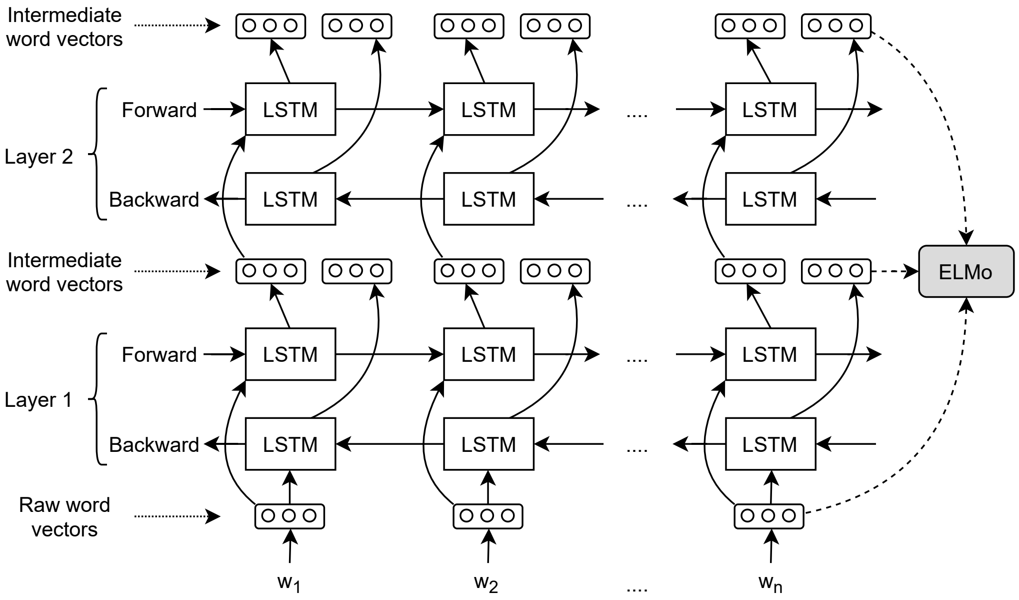

Figure 2: ELMo architecture.

vocabulary. ELMo embeddings are trained on the basis of character sequences creating an ability to learn spelling variations as well as morphology of the languages.

The Figure 2 shows the ELMo architecture which contains three neural network layers. First layer is a convolutional layer which operates on character sequences. The second and third layers are bidirectional layers where each layer contains two concatenated LSTMs. The output of the third layer produces the contextualized word embeddings for the given text. The details of corpora which have been used to train ELMo embeddings for both Tamil and Tulu code-mixed datasets are given in the Table 3 and the Table 4 respectively. The vocabulary for both languages has been used with minimum frequency of 20. The ELMo projection dimension size was set to 256 and the ELMo model was run for five epochs for each language.

5 Results

The macro average metric has been used to evaluate the prediction of the shared task. The findings of the shared task are presented by Hegde et al. (2023). In our submissions, we achieved the macro average F_1-scores of 0.2150 and 0.5220 for Tamil and Tulu code-mixed datasets respectively. However, our initial models lacked English data, proper tokenization and *demojization*. We retrained the models and the updated results for development and test sets which are shown in the Tables 5, 6, 7 and 8.

Category	Pre.	Rec.	F_1-score
Positive	0.7669	0.7944	0.7804
Negative	0.3545	0.6521	0.4593
Mixed feelings	0.3699	0.1461	0.2095
Unknown state	0.5128	0.3290	0.4008
micro avg	0.6263	0.6263	0.6263
macro avg	**0.5010**	**0.4804**	**0.4625**
weighted avg	0.6277	0.6263	0.6124

Table 5: Sentiment analysis results for Tamil code-mixed development set.

Category	Pre.	Rec.	F_1-score
Positive	0.1235	0.4247	0.1914
Negative	0.6340	0.3639	0.4624
Mixed feelings	0.2404	0.2475	0.2439
Unknown state	0.3000	0.2190	0.2532
micro avg	0.3220	0.3220	0.3220
macro avg	**0.3245**	**0.3138**	**0.2877**
weighted avg	0.4448	0.3220	0.3537

Table 6: Sentiment analysis results for Tamil code-mixed test set.

The results for Tamil code-mixed sentiment analysis have been improved by enhancing the size and the quality of the datasets for transfer learning. The macro average F_1-score has been improved with a gain of 0.0727 points. The model performed better on development set with a macro average F_1-score of 0.4625 as compared to the test set. The reason behind the significant difference is the size of evaluation sets for Tamil. The development set contains 3,786 sentence whereas the test set contains only 649 sentences.

On the other hand, the Tulu evaluation sets have almost same number of samples. The Tulu model performed with the macro average F_1-scores of 0.5386 and 0.5133 on development and test set respectively. The Tulu dataset has less variations as compared to Tamil data as it mostly contains Romanized Tulu comments which resulted higher results.

Category	**Pre.**	**Rec.**	**F_1-score**
Positive	0.8506	0.7561	0.8006
Negative	0.5435	0.2778	0.3676
Neutral	0.5096	0.7921	0.6202
Mixed feelings	0.4194	0.3250	0.3662
micro avg	0.6440	0.6440	0.6440
macro avg	**0.5807**	**0.5377**	**0.5386**
weighted avg	0.6607	0.6440	0.6373

Table 7: Sentiment analysis results for Tulu code-mixed development set.

Category	**Pre.**	**Rec.**	**F_1-score**
Positive	0.8615	0.7413	0.7969
Negative	0.5588	0.3167	0.4043
Neutral	0.5034	0.7614	0.6061
Mixed feelings	0.2875	0.2150	0.2460
micro avg	0.6314	0.6314	0.6314
macro avg	**0.5528**	**0.5086**	**0.5133**
weighted avg	0.6494	0.6314	0.6273

Table 8: Sentiment analysis results for Tulu code-mixed test set.

From the results, it is quite evident that transfer learning has the ability to produce competitive results for code-mixed corpora. However, the data preparation is an important task before the training process. There should also be a balance in the corpus having representation of various types of comments. In this way, the sentiment analysis models could perform better. The sentiment analysis of code-mixed text is important research topic which requires more research to analyze online text.

6 Conclusion

This paper presents the sentiment analysis of code-mixed Tamil and Tulu YouTube comments. The code-mixed text contains text from different scripts, such as, Tamil, Tulu, English, Romanized Tamil and Tulu, mixed text and emoticons. The variety of languages makes it quite challenging to predict sentiments with higher accuracy. We proposed a Bidirectional Long-Short Term Memory networks based model for both languages which further uses contextualized word embeddings at the input layers of the model. For that purpose, ELMo embeddings have been trained on larger unannotated code-mixed text corpora. The transfer learning by using trained ELMo models for both language was quite helpful to achieve improved sentiment analysis results. Our models performed with the macro average F_1-scores of 0.2877 and 0.5133 on Tamil and Tulu code-mixed datasets respectively.

References

Gazi Imtiyaz Ahmad, Jimmy Singla, Ali Anis, Aijaz Ahmad Reshi, and Anas A Salameh. 2022. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus-a comprehensive review. *International Journal of Advanced Computer Science and Applications*, 13(2).

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Bharathi Raja Chakravarthi, Mihael Arcan, and John Philip McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020a. 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. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Forum for information retrieval evaluation*, pages 21–24.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-Lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

Toqeer Ehsan and Miriam Butt. 2020. Dependency Parsing for Urdu: Resources, Conversions and Learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5202–5207.

Toqeer Ehsan and Sarmad Hussain. 2020. Development and Evaluation of an Urdu Treebank (CLE-UTB) and a Statistical Parser. *Language Resources and Evaluation*, pages 1–40.

Toqeer Ehsan and Sarmad Hussain. 2022. *Statistical Parser for Urdu*. Ph.D. dissertation, University of Engineering and Technology, Lahore, Pakistan.

Toqeer Ehsan, Javairia Khalid, Saadia Ambreen, Asad Mustafa, and Sarmad Hussain. 2022. Improving Phrase Chunking by using Contextualized Word Embeddings for a Morphologically Rich Language. *Arabian Journal for Science and Engineering*, pages 1–19.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022a. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in Tamil and Tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Anand Kumar Madasamy, and Bharathi Raja Chakravarthi. 2022b. A Study of Machine Translation Models for Kannada-Tulu. In *Congress on Intelligent Systems*, pages 145–161. Springer.

Prashanth Kannadaguli. 2021. A code-diverse tulu-english dataset for nlp based sentiment analysis applications. In *2021 Advanced Communication Technologies and Signal Processing (ACTS)*, pages 1–6. IEEE.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. *arXiv preprint arXiv:1802.05365*.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech & Language*, 76:101407.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

Amina Tehseen, Toqeer Ehsan, Hannan Bin Liaqat, Amjad Ali, and Ala Al-Fuqaha. 2022. Neural POS Tagging of Shahmukhi by Using Contextualized Word Representations. *Journal of King Saud University-Computer and Information Sciences*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020. Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts. In *2020 Moratuwa engineering research conference (MERCon)*, pages 272–276. IEEE.

Aditya Vyawahare, Rahul Tangsali, Aditya Mandke, Onkar Litake, and Dipali Kadam. 2022. PICT@ DravidianLangTech-ACL2022: Neural Machine Mranslation on Dravidian Languages. *arXiv preprint arXiv:2204.09098*.