Habesha@DravidianLangTech: Utilizing Deep and Transfer Learning Approaches for Sentiment Analysis

Mesay Gemeda Yigezu[1], **Tadesse Kebede**[2] , **Olga Kolesnikova** [3],
Grigori Sidorov [4], **Alexander Gelbukh** [5]
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico [1], [3], [4], [5]
Haramaya University, Ethiopia [2]
Correspondence: mgemedak2022@cic.ipn.mx

Abstract

This research paper focuses on sentiment analysis of Tamil and Tulu texts using a BERT model and an RNN model. The BERT model, which was pretrained, achieved satisfactory performance for the Tulu language, with a Macro F1-score of 0.352. On the other hand, the RNN model showed good performance for Tamil language sentiment analysis, obtaining a Macro F1-score of 0.208. As future work, the researchers aim to fine-tune the models to further improve their results after the training process.

1 Introduction

The recent rise of the internet as websites, blogs, social networks, online portals, and content-sharing services contribute to the huge amount of user-generated sentiment texts Kebede (2019). These sentiment texts are very important to get feedback from people. Thus, the sentiment analysis fields that help to automatically examine sentiment from texts are required in various areas. Sentiment analysis is a task of subjectivity analysis with some linkage to effective computing principles (Pang et al., 2008). Sentiment analysis is an interdisciplinary theme that exploits natural language processing (NLP) methods, text mining, and computational linguistics to recognize and extract subjective information from source materials (Khan et al., 2009). It involves classifying the sentiment data into different polarity levels such as positive, negative, and neutral. Sentiment analysis is used by the sectors such as advertisers, movie creators, booksellers, political parties, supermarkets, industries, restaurants, etc. that demand their customers' feedback on a particular issue to improve themselves afterward. For example, (1), sentiment analysis is used in government elections to examine the people's sentiments, appraisals, attitudes, and emotions on the candidates of the election. (2), business companies introduce new products to the market and then usefully extract the sentiment of the people on the internet about the product, and settle the future feasibility of this new product. It is possible to gather the sentiment from customers using surveys, blogs, and suggestion committees but this may result in a waste of resources and not provide us the comment in a short period. Automatically gathering and classifying the sentiment that is expressed by natural language on social networks eradicates those problems with minimum time and resources. The objective of this work is to develop a sentiment analysis for Tamil and Tulu languages adopting the pre-trained model and Deep Learning algorithm and to evaluate the model's performance. Getting the sentiment of people in the business sector organizations engaged leads to success in the field or business(Kebede, 2019). The rest of the study is organized as follows. Currently, opinions of the people towards any organization in the world are collected by holding different conferences/meetings by using oral and manual collection methods. Such ways of gathering people's feedback consume much time and resources. In order to solve this, knowing what people are interacting with in their day-to-day life helps us to collect sentiments regarding some issues in a little time. Nowadays, the internet technology that is invented in the world like Facebook, Twitter, Blogs, Websites, etc. is usable in using Tulu and Tamil language. So, the Tamil and Tulu language usage on internet technology provides us with massive sentiment data. Even though the Tamil and Tulu language sentiment data is generated by different organizers, there is no sufficient sentiment analysis task that classifies the sentiment text into their polarity level. Thus, collecting this sentiment and classifying it into their polarity level is the intention of this work. This may optimize oral and survey-based sentiment collection by adopting pre-trained and deep-learning algorithms. Thus, the aim of this research is:

- To find the best data processing techniques for Tulu and Tamil sentiment data.

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 239–243, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_035

- To investigate techniques that should be applied to perform Tamil and Tulu language sentiment classification into their polarity level.
- To evaluate the Tamil and Tulu sentiment analyzer model using Macro F1-score.

2 Related Work

In this part, related sentiment analysis research investigating Tamil or Tulu languages using different approaches is presented. Besides, the recent sentiment analysis task which is developed by state of art methods is reviewed. (Thavareesan and Mahesan, 2018) performed a review to analyze the recent literature in the area of sentiment analysis in Tamil texts. This study considered the preprocessing, corpus and techniques, and success rate for review. The paper finds that the performance of the model relies on the preprocessing steps such as negation handling and stop word removal. The review concluded that SVM and RNN classifiers taking Word level feature representation using TF-IDF and Word2vec provide better performance than grammar rules-based classifications and other classifiers with usage of words, TF, and BoW features. The review revealed that resources such as Tamil SentiWordNet, adjective rules, and n-grams also can be used in SA on Tamil text as it proves a significant performance.

To analyze the sentiment content of Tamil Movie reviews sentiment analysis task is investigated by (Ramanathan et al., 2021). The aim of this work was to categorize the sentiment of Tamil movies based on Tamil tweets using Tamil SentiWordNet (TSWN). Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction methods are applied to identify the sentiment polarity of the Tamil movie data set. In this work, Tamil SentiWordNet is used with adjectives to classify the sentiment. Since the adjectives are not sufficient to detect the sentiment in the Tamil texts as future work, the paper intended to use adverbs in the Tamil language to identify the sentiment. Finally, the paper recommended the adjective wordnet for any language is appropriate for the sentiment analysis.

This research (Roy and Kumar, 2021) developed the sentiment analysis on Tamil code-mixed text using Bi-LSTM. The code-mixed data like Hindi-English, Malayalam-English, and similar ones; are used to detect the sentiment from Tamil review data. The proposed Bi-LSTM framework automatically extracts the features from input sentences and predicts their sentiment with a 0.552 F1 score for the best case using. (Divyasri et al., 2022) studied emotion analysis of Tamil texts using Language Agnostic Embeddings. The paper invented a multilingual transformer model for the emotion analysis of Tamil text as required by the DravidianTechLang ACL 2022 shared task. In this paper, LaBSE, a pre-trained language agnostic BERT model, was found to perform comparatively well on the Tamil dataset. The paper claimed the result of Tamil Sentiment Analysis can be optimized by utilizing custom embeddings, based on a statistical analysis of the language, to process the data before training the model. A code-diverse Tulu-English data for NLP-based sentiment analysis applications are developed by this paper (Kannadaguli, 2021). The research concentrated on NLP for emotion and sentiment detection of Tulu, a vibrant South Indian language, to start with. Development of the standard corpus for NLP applications of code-diverse text in Tulu-English is contributed by this work. The performance analysis of the dataset is performed through Krippendorff's Alpha value of 0.9 indicates that it is a benchmark in the development of the Automatic Sentiment Analysis system for Tulu. Though research (Kurniasari and Setyanto, 2020), (Topbaş et al., 2021) are conducted for different languages using deep learning and transformer learning the Tulu and Tamil sentiment analysis is very limited using these methods.

3 Task description

To identify the sentiment polarity of the code-mixed dataset of comments and posts in Tamil-English and Tulu-English collected from social media by DravidianLangTech2023 (Chakravarthi et al., 2020) is used. The comment/post may contain more than one sentence but the average sentence length of the corpora is 1. Each comment or post is annotated with sentiment polarity at the comment or post level. This dataset also has class imbalance problems depicting real-world scenarios. Our proposal aims to encourage research that will reveal how sentiment is expressed in code-mixed scenarios on social media. The purpose of this task is to perform sentiment analysis for Tulu and Tamil languages. The given Tulu sentiment text is classified into mixed feelings, negative, neutral, and positive. On the other hand, the Tamil sentiment data is classified into classes such as negative, mixed feelings, positive and unknown (Hedge et al., 2023).

The given data is full of noise and to upgrade the performance of the model the data cleaning task is performed at the pre-processing stage for both languages. The participants were provided with the Tamil and Tulu development, training, and test dataset. This is a message-level polarity classification task. Given a Youtube comment, systems have to classify it into positive, negative, neutral, or mixed emotions. The participants will be provided development, training, and test dataset code-mixed text in Dravidian languages (Tamil-English and Tulu-English).

4 Methodology

(Roy and Kumar, 2021) (Divyasri et al., 2022) (Kurniasari and Setyanto, 2020) (Topbaş et al., 2021) (Talaat, 2023) (Zhang et al., 2023) researchers suggested the pre-trained model like BERT and deep learning algorithms perform great for the sentiment analysis. The review paper (Cui et al., 2023) recommends the deep learning technology, with its hybrid methods combining sentiment dictionary and semantic analysis, fine-grained sentiment analysis methods, and non-English language analysis methods, and cross-domain sentiment analysis techniques have gradually become the research trends. The overall proposed framework of Tamil and Tulu sentiment analysis is depicted in Figure 1.

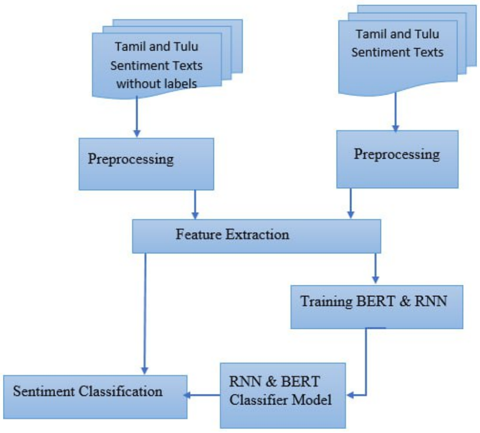
Tamil and Tulu Sentiment Texts without labels
Tamil and Tulu Sentiment Texts
Preprocessing
Preprocessing
Feature Extraction
Training BERT & RNN
Sentiment Classification
RNN & BERT Classifier Model

Figure 1: Proposed architecture for sentiment analysis

The proposed framework incorporates data preprocessing, feature extraction, training a model, and sentiment classification evaluation. The dataset used in this study is found on DravidianLangTech2023. Majorly the dataset of the Tamil language belongs to the positive sentiment class. The rest of the Tamil data set is distributed to Negative, mixed feelings, and unknown state sentiment classes. Likewise, the Tulu sentiment is majorly included as a positive sentiment class and the others are distributed as Negative, Neutral, and Mixed feeling sentiment categories. Figure 2 shows the distribution of the Tamil and Tulu sentiment analysis data.

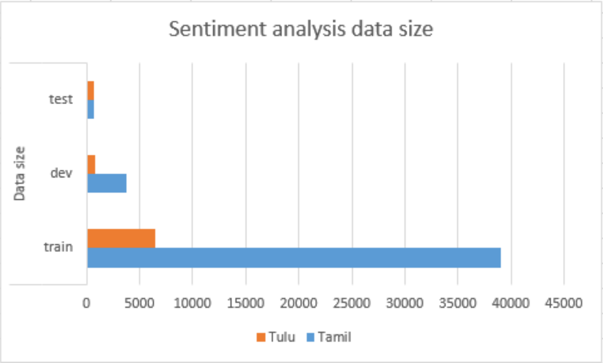

Figure 2: distribution of data size for sentiment analysis

Based on the provided figure, it is evident that there is a limited quantity of both development and training data. The only exception is the training data for the Tamil language, although even these datasets are unbalanced. As a result, this situation causes the model to exhibit reduced or inaccurate performance.

4.1 Data pre-processing

In the data preprocessing stage the unnecessary noises are removed from the original Tamil and Tulu sentiment data. The pre-processing stage removes HTML tags, URLs, digits, and punctuation marks, and lowercasing is performed. In addition to this, the emoji is replaced by the text in both Tamil and Tulu sentiment data.

4.2 Feature extraction

The feature is extracted from the processed data before it can be used in a model. First, for the RNN model, TextVectorization and embedding are used to extract the feature. During embedding the one vector per word is stored. Secondly, before training the BERT to convert our Tulu and Tamil sentiment data into numerical values methods such as Bag of words, TFIDF, and Word Embedding are used to extract the features. In this task, the Tokenizer class from pre-trained DistilBert is used in order to tokenize the Tamil and Tulu sentiment data.

4.3 Model Training

In this work to generate the Tulu and Tamil sentiment analysis model a pre-trained BERT model and deep learning algorithm: RNN is used. BERT is a machine learning technique developed by Google based on the Transformers mechanism. Sometimes, BERT models are used in place of conventional RNN based because the BERT model suffered from information loss in large sequential text. While the BERT can easily understand the context of a word in a sentence based on previous words in the sentences due to its bi-directional approach. RNNs are a form of Artificial Neural networks that can memorize arbitrary-length sequences of input patterns by capturing connections between sequential data types. A neural network that is intentionally run multiple times, where parts of each run feed into the next run. Specifically, hidden layers from the previous run provide part of the input to the same hidden layer in the next run. Recurrent neural networks are particularly useful for evaluating sequences so that the hidden layers can learn from previous runs of the neural network on earlier parts of the sequence.

4.4 Performance metrics

To evaluate the model performance, the classification metrics called F1-score are used. The classification system's performance will be measured in terms of macro-averaged Precision, macro-averaged Recall, and macro-averaged F-Score across all the classes. The macro average F1-score is the harmonic mean of the precision and recall (Pang et al., 2008). Precision is defined as the number of correctly predicted sentiment categories among the retrieved instances of the particular sentiment category. The recall is defined as the number of correctly predicted sentiment categories among the total number of instances of that particular sentiment category.

5 Experiments and Results

5.1 Experimental

The experimentation is done on the Python Tensorflow which is a popular and widely used machine learning framework for developing deep learning applications. Installing Libraries and dependencies make ready all the necessary libraries and packages. To ensure that the GPU is enabled, the TensorFlow API: 'tf. config' is used. In this article, the BERT and RNN are used for developing a sentiment analysis for Tamil and Tulu Sentiment Analysis. BERT is trained on 3 epochs, 16 batch sizes, and using a 5e-5 Learning rate (Adam). The RNN is trained using parameters: buffer size: 10000, batch size: 8, and vocab size:10000. Finally, Both BERT and RNN model is developed for both Tulu and Tamil sentiment classification.

5.2 Results

In this task, we assess the effectiveness of development models for sentiment analysis in two different languages, Tamil and Tulu. The evaluation of these models is based on the Macro F1-Score metric, which provides an overall measure of performance across multiple classes.

Upon analyzing the test dataset, the results reveal that the BERT model outperformed other models in the Tulu language, achieving a Macro F1-Score of 0.35. On the other hand, for sentiment analysis in the Tamil language, the developed RNN model exhibited superior performance, attaining a Macro F1-Score of 0.20.

These findings highlight the suitability of the BERT model for sentiment analysis in the Tulu language, indicating its capability to capture and understand the nuances of sentiment expressed in Tulu text. Conversely, the RNN model demonstrates its proficiency in capturing the sentiment complexities of the Tamil language, making it the preferred choice for sentiment analysis in Tamil. Table 1 depicts a summary of the result.

Tasks	Macro-score			
	P	R	F1	Acc
Tamil	0.25	0.21	0.20	0.40
Tulu	0.36	0.35	0.35	0.57

Table 1: Experimental result

6 Conclusion

In this research paper, we have trained a BERT and RNN model for the sentiment analysis of Tamil and Tulu texts as the need of DravidianLangTech 2023-RANLP 2023. A pre-trained BERT language BERT model performed well for the Tulu language comparing it with the Tulu language yielding a Macro F1 score of 0.352. In another way, the RNN model performance is good for Tamil language sentiment analysis with a Macro F1 score of 0.208. The researchers take as future work fine-tuning the models to improve the results of the models after training.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America Ph.D. Award.

References

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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.

Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, pages 1–42.

K Divyasri, GL Gayathri, Krithika Swaminathan, Thenmozhi Durairaj, B Bharathi, et al. 2022. Pandas@ tamilnlp-acl2022: Emotion analysis in tamil text using language agnostic embeddings. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 105–111.

Asha Hedge, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya S.K, Durairaj Thenmozhi, Martha Karunakar, Shreya Sriram, 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.

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.

Tadesse Kebede. 2019. *MACHINE LEARNING BASED MULTI-SCALE SENTIMENT ANALYSIS FOR AFAAN OROMO POSTS*. Ph.D. thesis, Haramaya university.

Khairullah Khan, Baharum B Baharudin, Aurangzeb Khan, et al. 2009. Mining opinion from text documents: A survey. In *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies*, pages 217–222. IEEE.

Lilis Kurniasari and Arif Setyanto. 2020. Sentiment analysis using recurrent neural network. In *Journal of Physics: Conference Series*, volume 1471, page 012018. IOP Publishing.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Vallikannu Ramanathan, T Meyyappan, and SM Thamarai. 2021. Sentiment analysis: An approach for analysing tamil movie reviews using tamil tweets. *Recent Advances in Mathematical Research and Computer Science*, 3:28–39.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment analysis on tamil code-mixed text using bi-lstm. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Amira Samy Talaat. 2023. Sentiment analysis classification system using hybrid bert models. *Journal of Big Data*, 10(1):1–18.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2018. Review on sentiment analysis in tamil texts.

Ayşenur Topbaş, Akhtar Jamil, Alaa Ali Hameed, Syed Muzafar Ali, Sibghatullah Bazai, and Syed Attique Shah. 2021. Sentiment analysis for covid-19 tweets using recurrent neural network (rnn) and bidirectional encoder representations (bert) models. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pages 1–6. IEEE.

Xiangsen Zhang, Zhongqiang Wu, Ke Liu, Zengshun Zhao, Jinhao Wang, and Chengqin Wu. 2023. Text sentiment classification based on bert embedding and sliced multi-head self-attention bi-gru. *Sensors*, 23(3):1481.