# Sentiment Analysis using Cross Lingual Word Embedding Model

C Jerin Mahibha[1], Sampath Kayalvizhi[2] and Durairaj Thenmozhi[3]

[1]*Meenakshi Sundararajan Engineering College, Chennai*
[2]*Sri Sivasubramania Nadar College of Engineering, Kalavakkam*
[3]*Sri Sivasubramania Nadar College of Engineering, Kalavakkam*

**Abstract**

Sentiment analysis deals with analysing the given text and classifying whether the text is a positive, negative or neutral one. The sentiment analysis forms the base for applications where the public views could be known. This paper shows how a multi label classification of the given text could be implemented by considering the sentiment associated with the text. The models that are applied for monolingual sentiment analysis may not provide good results when it is extended for code mixed data. When a Cross Lingual model was applied to the training data-set provided by Dravidian Code-mix FIRE 2021 for the Task A which uses Tamil - English code mixed data, it was able to classify the test data-set with an average F1 score of 0.514. .

**Keywords**

Sentiment analysis, Code-mixed data, Transformer model, Imbalanced data-set, Sampling

## 1. Introduction

India is a land of large bilingual communities. Most part of the Indian community are well versed in the language English [1] in addition to their native language. So when Indians express their opinions considering various situations as tweets or comments, mostly it is a mix of English and a regional language which is represented as code mixed data. Hence analysis of code mixed data has become an important part of any analysis considering the Indian society.

Sentiment analysis is a research area under Natural language Processing where the sentiment associated with a text has to be identified [2]. Usually the sentiment analysis process is expressed as a multi label classification problem[3] with a minimum of three sentiments associated with a text which can be represented as positive, negative or neutral. Sentiment expresses the inner feeling of a person towards a particular situation which can also be represented as emotions [4] towards a situation which can be a product, idea, concept, event etc.

When a corpus is created for a specific purpose like Sentiment analysis, the data in the corpus may not be uniformly distributed. It may have a particular class of data much more than

*FIRE 2021 : Forum for Information Retrieval Evaluation, December 13-17, 2021, India*

jerinmahibha@gmail.com (C. J. Mahibha); kayalvizhis@ssn.edu.in (S. Kayalvizhi); d_theni@ssn.edu.in (D. Thenmozhi)

other classes or a particular class of data may be much smaller than the other classes[5]. The sentiments associated with such imbalanced code-mixed data could be analysed using sampling techniques [6]. Various machine learning approaches namely, Random Forest Classifier, Logistic Regression, XGBoost classifier, Support Vector Machine and Naïve Bayes Classifier can be applied over this set to perform the classification process associated with the sentiment of the text. Lexicon based approaches can also be used for the classification process [7] which mainly rely on lexical resources like Wordnet which represents the words and their associated sentiment in order to perform the classification.

During this pandemic period, exposure to social media has seen a wide increase. Due to the use of online resources and online media specifying their views and posting comments in Social media sites have become very common among people. Indians prefer to express themselves using code mixed languages. These are difficult to analyse as they may not be grammatically formed with variations in spelling and use of abbreviations and also it has limited resources [8]. Deep learning models could be effectively utilized for performing the multi label classification of sentiment analysis over the code mixed data. Cross-lingual pretrained transformer models are expected to provide better performance for such task when it is associated with low-resourced languages [9].

## 2. Related works

In a code mixed corpus, usage of both code-mixed and non-code-mixed data are common which had been identified by creation of a lexicon dictionary [6] for the code-mixed corpus using which the problems associated with spelling variations, abbreviation usage had been handled and then the machine learning techniques had been applied for classification based on the sentiment associated with the text. Problems associated with the imbalance nature of the corpus had been handled by applying various sampling techniques over the corpus.

Sentiment analysis of Kannada-English code mixed corpus which had been created by crawling Facebook comments and the performance of various machine learning and deep learning techniques over the corpus using a distributed representation had been demonstrated by [10]. Sentiment analysis of a Hindi data-set had been implemented using the concept of cross-lingual contextual word embeddings and zero-shot transfer learning to project the predictions from resource-rich English to resource-poor Hindi language [11]. Classification of code mixed Hindi text based on sentiment had been implemented with the use of TF-IDF feature vectors of character n-grams where n ranged from 2 to 6 with an ensembled voting classifier and linear SVM classifier[12].

Polarity of Dravidian code-mixed comments had been identified using a sub-word level model and a word embedding based model which in turn had made use of Long Short Term Memory (LSTM) network and a machine learning based architecture which had used Inverse document frequency (TF-IDF) vectorization along with a Logistic Regression model [13] for the

**Table 1**
Data set

| Category | Training Set | Validation Set |
|---|---|---|
| Positive | 4271 | 480 |
| Negative | 5628 | 611 |
| Mixed Feeling | 4021 | 438 |
| Not Known | 20069 | 2257 |
| Not Tamil | 1667 | 176 |

classification task. The Decision Tree Algorithm had been used to computationally identify and categorize the opinions expressed as text in Kannada language [14].

To analyse code mixed data, bilingual embedding techniques has to be replaced with multilingual word embedding schemes [15] to achieve improvement in performance in any application associated with code-mixed data. For implementing sentiment analysis of code-mixed data, using different kinds of multilingual and cross-lingual embeddings, knowledge can be efficiently transferred from monolingual text to code-mixed text [16]. Variations of code mixed words had been captured by a cluster based preprocessing approach and then the sentences of code-mixed and standard languages had been mapped to a common sentiment space for performing the sentiment analysis by [17].

## 3. Data set Description

The data-set provided for the shared task Dravidian Code-Mix FIRE 2021 [2],[3] was a new gold standard corpus for sentiment analysis of code-mixed data. The text represents comment / post with an average sentence length of the corpora as 1. Each text is annotated with sentiment polarity at the comment / post level. The data-set had three sets of data, one for training, second for validation and the third set of data for testing based on which the results of the shared task was announced. The complete data-set had five different classes to which the data belong, namely Not-known, Positive, Negative, Not-Tamil and Mixed Feelings.

The table 1 shows the number of instances under each category both in the training and the validation data set. Considering the distribution of data in the data set it showed an imbalanced nature of the real world scenario with 56% of the data falling under Not-Known category in both training and validation sets. The instances of the test data-set has to be placed under any one of the above categories which was done by predictions based on the proposed model. The average F1 score of this predicted class of the test data was used by Dravidian Code-Mix FIRE 2021 to evaluate the proposed model and based on the predicted values, average F1 score was computed and the models were ranked.

The Fig:1 shows the distribution of the data in both the training and validation data-set provided by Dravidian Code-mix Fire 2021 for the Task1. It was evident that the data-set is an

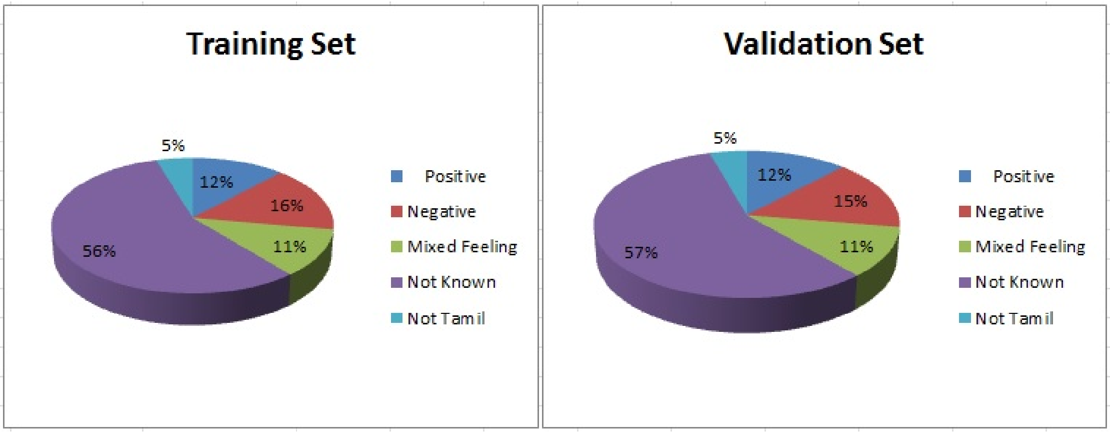


**Figure 1:** Distribution of data in the data set

imbalanced one with more instances under the category Not- Known.

As separate data-sets were provided for training and validation purposes, they were used for training and validating the proposed model. A separate test data-set was provided for which the sentiments were predicted using the trained model which were used for computing the F1 score of the model.

## 4. Proposed methodology

The shared task Dravidian Code-Mix FIRE 2021 was a multi-label classification task for sentiment analysis of Dravidian Code-mixed data based on the provided gold standard corpus[**?** ]. We as a team participated in the task associated with the sentiment analysis of Tamil-English code mixed text which had comments from YouTube. We have implemented the classification using a Cross Lingual pre-trained model.

The given data-set was imbalanced in nature, as a result of which the model was not able to turn up with good results. The data from the training data-set was down-sampled to enforce a balanced nature to the data-set. Using the Cross Lingual model substantial improvement in performance of the model is expected over both low-resource languages and high-resource languages. The Cross Lingual Model is trained with a Translation Language Modeling which helps the model to learn similar representations for different languages[1]. The vocabulary supported by the model was the default value which is 30145, and it uses 2048 encoder and pooling layers and 12 hidden layers with a dropout value of 0.1. The model supports language embedding with two languages supported by the model as the considered data set has text in code-mixed Tamil and English. The model was trained using the training data set and validated

[1] https://huggingface.co/transformers/model_doc/xlmroberta.html

**Table 2**
Submission score

| Parameters | Score |
|---|---|
| Precision | 0.615 |
| Recall | 0.485 |
| F1 Score | 0.514 |

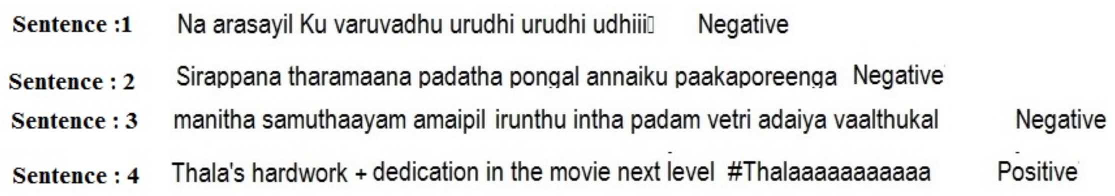
using the validation data set provided for the shared task by Dravidian Code-Mix FIRE 2021. The evaluation of the model was based on a separate test data set.

The performance of Dravidian Code-mix FIRE 2021 was evaluated based on the performance of the proposed system which in turn was measured in terms of weighted averaged precision, weighted averaged recall and weighted averaged F-Score across all the classes. Our model was able to provide an average F1 score of 0.514. Table 2 shows the value of various performance measures for sentiment predictions done for the text in the test data-set using the proposed model.

## 5. Error Analysis

The performance measures of the model shows that there is much more scope to improve the accuracy of the sentiment analysis of code mixed Tamil-English text. Imbalance nature of the data-set could be considered as one possible reason for this output which could be balanced by using up-sampling mechanisms instead of down-sampling the data. The down-sampled data could have made the model to miss out few key features which would have led to miss-classification of the text from the test data-set. Pre-processing of the data is equally important as up-sampling of the system as the posts retrieved from YouTube would have symbols and abbreviations.

Figure : 2 shows few example sentences which has not been correctly predicted by the proposed model. First three sentences have been identified to belong to the Negative category, but analysing the words in the sentences shows that they express positive sentiments. Sentence 4 had been classified as positive by considering the presence of the word 'level', but considering the sentence as a whole it should be classified as Not-Tamil. Fine tuning the proposed model and avoiding the loss of information would help to reduce the miss-classifications that have occurred in the model.



Sentence :1 Na arasayil Ku varuvadhu urudhi urudhi udhiii Negative

Sentence : 2 Sirappana tharamaana padatha pongal annaiku paakaporeenga Negative

Sentence : 3 manitha samuthaayam amaipil irunthu intha padam vetri adaiya vaalthukal Negative

Sentence : 4 Thala's hardwork + dedication in the movie next level #Thalaaaaaaaaaa Positive

**Figure 2:** Sample Sentences

## 6. Conclusion

From the result it could be found that the performance achieved by our model on the new golden corpus provided by Dravidian Code-mix FIRE 2021 for code mixed Tamil-English text could be improved. As numerous multilingual based transformer models and approaches are available, more scope is there for research to be carried out in this field to identify a better model for sentiment analysis of code mixed text.

## References

[1] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus creation for sentiment analysis in code-mixed tamil-english text, arXiv preprint arXiv:2006.00206 (2020).

[2] B. R. Chakravarthi, R. Priyadharshini, S. Thavareesan, D. Chinnappa, D. Thenmozhi, E. Sherly, J. P. McCrae, A. Hande, R. Ponnusamy, S. Banerjee, C. Vasantharajan, Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[3] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, D. Thenmozhi, E. Sherly, Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada, in: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, 2021.

[4] A. Kalaivani, D. Thenmozhi, SSN_NLP_MLRG@ dravidian-codemix-FIRE2020: Sentiment code-mixed text classification in tamil and malayalam using ulmfit., in: FIRE (Working Notes), 2020, pp. 528–534.

[5] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://aclanthology.org/2020.sltu-1.28.

[6] R. Srinivasan, C. Subalalitha, Sentimental analysis from imbalanced code-mixed data using machine learning approaches, Distributed and Parallel Databases (2021) 1–16.

[7] N. H. Mahadzir, et al., Sentiment analysis of code-mixed text: A review, Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (2021) 2469–2478.

[8] S. R. Shah, A. Kaushik, Sentiment analysis on indian indigenous languages: a review on multilingual opinion mining, arXiv preprint arXiv:1911.12848 (2019).

[9] G. Lample, A. Conneau, Cross-lingual language model pretraining, arXiv preprint arXiv:1901.07291 (2019).

[10] K. Shalini, H. B. Ganesh, M. A. Kumar, K. P. Soman, Sentiment analysis for code-mixed indian social media text with distributed representation, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1126–1131. doi:10.1109/ICACCI.2018.8554835.

[11] A. Kumar, V. H. C. Albuquerque, Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language, Transactions on Asian and Low-Resource Language Information Processing 20 (2021) 1–13.

[12] P. Mishra, P. Danda, P. Dhakras, Code-mixed sentiment analysis using machine learning and neural network approaches, arXiv preprint arXiv:1808.03299 (2018).

[13] A. V. Mandalam, Y. Sharma, Sentiment analysis of Dravidian code mixed data, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 46–54. URL: https://aclanthology.org/2021.dravidianlangtech-1.6.

[14] P. Ranjitha, K. Bhanu, Improved sentiment analysis for dravidian language-kannada using dicision tree algorithm with efficient data dictionary, in: IOP Conference Series: Materials Science and Engineering, volume 1123, IOP Publishing, 2021, p. 012039.

[15] A. Pratapa, M. Choudhury, S. Sitaram, Word embeddings for code-mixed language processing, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 3067–3072.

[16] S. Yadav, T. Chakraborty, Unsupervised sentiment analysis for code-mixed data, arXiv preprint arXiv:2001.11384 (2020).

[17] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of code-mixed languages leveraging resource rich languages, arXiv preprint arXiv:1804.00806 (2018).