SA-SVG@Dravidian-CodeMix-FIRE2020: Deep Learning Based Sentiment Analysis in Code-mixed Tamil-English Text

Anbukkarasi S[a], Varadhaganapathy S[b]

[a]*Velalar College of Engineering and Technology), Erode, Tamilnadu*
[b]*Kongu Engineering College, Erode, Tamilnadu*

Abstract

Sentiment Analysis (SA) is the process of identifying the opinions and thoughts on particular context. In recent times, People tend to share their feelings, emotions and ideas through social media applications such as Facebook, Twitter, Instagram. The bright side of these kinds of applications is users can interact with other people in code-mixed text. Analysing sentiments in code-mixed data is little bit tedious when compared in mono lingual texts as the code-switching increases the complexity. For conducting the experiments, the dataset given by Dravidian-CodeMix-FIRE2020 contains the code-mixed data of youtube comments has been used. Deep Learning based Bi-LSTM model is used for classification in our implementation. F1-Score, Precision ,Recall metrics are used for evaluation purpose. Our code is published in github at https://github.com/AnbukkarasiS/Dravadian-Codemix.

Keywords

Sentiment Analysis, Bi-LSTM, Code-Mixed, Tamil, Dravidian languages

1. Introduction

Sentiment Analysis is the emerging topic of the Natural Language Processing domain. It is the process of identifying the emotions like happy, sad, and depressed from the given texts. This analysis can be helpful in product review, movies reviews etc. It is helpful in making a decision on a particular task like movie watching, purchasing a product, book review. In countries like India, people speak different languages in different regions. Even within a state they speak different languages. Emergence of social media applications like Facebook, Twitter, Youtube makes the sentiment analysis task a little bit tricky as many of the users often use code-mixed text to opine their views. Code-mixed text consist of text in a mixed language, that too not written in native script. Since these texts are not in native script, classifying them into a particular class is a little bit tedious task [1, 2, 3, 4]. A sentence might be fully written in the roman script, or it might be a mixture of the languages.

Hence it is required to do the SA task in code-mixed data carefully. In this paper we proposed a deep learning based model for sentiment analysis in code-mixed data. The given text is classified as positive, negative, neutral, mixed-feeling and non-Tamil for Tamil and Malayalam

FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India
anbu.1318@gmail.com (A. S); varadhaganapathy@gmail.com (V. S)
0000-0003-0226-8150 (A. S)

Table 1
Test and train data statistics for Tamil-English code-mixed dataset

Data	Sentence Count
Train data	11335
Test data	3149

languages. Tamil and Malayalam are closely related languages from Dravidian language family [5]. The data set is provided by the shared task Dravidian-CodeMix-FIRE2020. The dataset consist of 15744 youtube comments, 11,335 training data, 1,260 validation data and 3,149 test data for code-mixed Tamil-English text. [6, 7]. The sample data with rough translation in English is given below:

- Positive: Sema thalaiva endrum ungalukku visvasamana fans thaan (Super hero, always your loyal fans)
- Negative: nanum vjs fan than ..enaku pudikala.. trailer (I am too vjs fan. But I don't like trailer)
- Neutral: Kaithiye ippadi irukuna,thalapathy 64 eppadi irukuna.VERITHANAMA IRUKUM (Even kaidhi looks good, so thalapathy 64 would be more rocking)
- Mixed-Feelings: Rajinikula expiry date mudinju over expiry akiruju (expiry date is over for Rajini) Title track ah vida trailer bgm sema.. yuvan (Trailer bgm is superb when compare to title track ..Yuvan)
- Non-Tamil: Seems like remake movie, amithab tapsee etc

In this paper, our motivation is to classify the given youtube comments as positive, negative, neutral, mixed-feeling and non-Tamil. The detail of the data set is given in Table 1.

This paper is classified as following. Section 2 describes the related work. Section 3 specifies the proposed methodology. Section 4 concludes the paper.

2. Related Work

There are very few works have been carried out for sentiment analysis in code-mixed data.

LSTM based sentiment analysis in tweets is performed in [8]. In this paper, authors use combined character based splitting for feeding the model Authors claim they achieved 86.2 % of accuracy when Bi-LSTM model is used and 77.2% of accuracy when LSTM is used. RNN based sentiment analysis is performed in micro texts in [9]. It is performed in the native script of the languages such as Tamil, Hindi and Bengali. This system achieved 88.23, 72.01, 65.16 % of accuracy for the languages Tamil, Hindi and Bengali respectively. It is said that unsupervised data could also be included in the model as a future work.

Vijay et.al created corpus and performed sentiment analysis in Hindi-English [10, 11] code-mixed social media text using Support Vector Machine (SVM) [12, 13]. Without including any

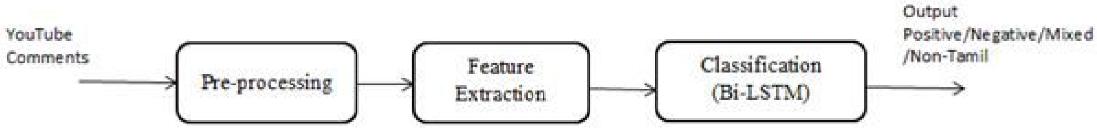

Figure 1: Proposed System - Bi-LSTM based Sentiment Analysis

features, they achieved 58.2% of accuracy. They claimed that character based classification increases the accuracy of the model to the significant amount. The system lacks in annotating the corpus on part-of-speech tag based. The size of the dataset is minimal in the work.

Shallow morphological parsers are used for analysing sentiments in online documents [14]. Based on the parsers, binary parse tree of recursive network is used. For long phrase sentences they achieved 71.1 % of accuracy. Corpus for Tamil-English code-mixed data has been created by [2]. This is the gold standard corpus for Tamil-English code-mixed data. For this work, various annotators label the classification of the given data. They have created a baseline model with Logistic regression, K-Nearest Neighbour, 1-dimensional convolution network etc.

A sentiment analysis dataset has been created for code-mixed Malayalam – English text by [4]. They have created the corpus with around 7743 sentences in code-mixed Malayalam-English text. For the baseline model, they used various machine learning approach such as Logistic regression, (LR), Support vector machine (SVM), Decision tree (DT), Random Forest (RF), Multinomial Naive Bayes (MNB), K-nearest neighbours (KNN).

In the proposed work, Bi-LSTM model has been used for sentiment analysis in code-mixed tamil-English text.

3. Proposed Methodology

This section describes the proposed methodology for classifying the given text as positive, negative, neutral, mixed-feeling and non-Tamil. Figure 1 depicts the overall system methodology proposed. In our proposed approach, a deep learning based Bi-LSTM model has been used for the classification purpose. Implementation has been done using the Tensorflow package.

- Pre-processing
- Feature Extraction
- Classification

The basic structure is given in Figure 1.

Pre-Processing

The input data contains some special symbols like ‘@’, ‘…’ etc. So the very first step is to clean the data. The symbols and special characters are removed from the given text and fed to the system.

Feature Extraction

Table 2
Parameters used in Bi-LSTM model

Parameter	Value
Optimizer	Adam
Activation	Sigmoid
Batch Size	128
Epoch	10
Loss Function	Sparse categorical cross entropy

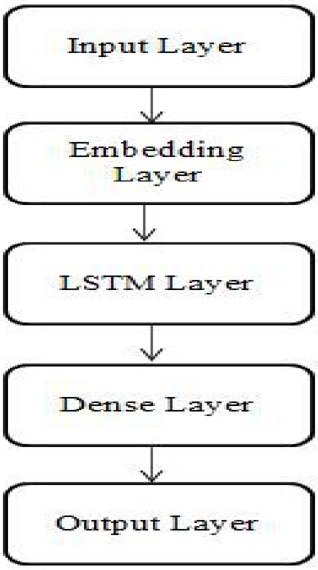

Figure 2: Bi-LSTM Model with layers

All the deep learning models accept the input in terms of numbers only. Text input cannot be fed to the models. Hence, the given code-mixed text is first converted into numerals. For this, the input sentence is broken into tokens. This part is known as tokenization. After this step, each token is mapped with the number vector. As each token is of different size, they have to be made equal by padding. The additional zeros at the right end of the given input sequence make the input data is of same size. Finally this input is fed to the Bi-LSTM model. It is implemented with Tensorflow package. The code for the same has been given in [15]

Classification

In this phase, the given input sequences are classified into the corresponding output. For classification, RNN based Long Short Term Memory model is used with Tensorflow package. The model outputs the given code-mixed youtube comments as Positive, Negative, Mixed-Feeling, Neutral and non-Tamil.

The various parameters used in the Bi-LSTM model are given in Table 2.

The detail of the Bi-LSTM model is represented in Figure 2.

Table 3
Bi-LSTM Model Performance. W- weighted average

Language	Dataset	W-Precision	W-Recall	W-F1 Score
Tamil	Validation	0.53	0.35	0.42
	Test	0.33	0.07	0.10

4. Results

For studying the performance of the given dataset, on Tamil-English code-mixed data, Bi-LSTM classifier is used with the parameters given in Table 2. We achieved the weighted average F-Score of 0.10 and ranked 14th in Tamil-English. The detailed results are given in Table 3.

5. Conclusion

The proposed system is used to classify the given input youtube comments into Positive, Negative, Mixed-Feeling, Neutral and non-Tamil on the code-mixed data given by Dravadian Codemix-FIRE 2020 task. The system uses Bi-LSTM model for the classification purpose. In future, the model could be improved by parameter tuning and other neural network models such as GRU with Attention mechanism could also used.

References

[1] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Wordnet gloss translation for under-resourced languages using multilingual neural machine translation, in: Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, 2019, pp. 1–7.

[2] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Comparison of different orthographies for machine translation of under-resourced Dravidian languages, in: 2nd Conference on Language, Data and Knowledge (LDK 2019), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[3] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[4] B. R. Chakravarthi, P. Rani, M. Arcan, J. P. McCrae, A survey of orthographic information in machine translation, arXiv preprint arXiv:2008.01391 (2020).

[5] B. R. Chakravarthi, N. Rajasekaran, M. Arcan, K. McGuinness, N. E.O'Connor, J. P. McCrae, Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages, in: Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects, Barcelona, Spain, 2020.

[6] B. R. Chakravarthi, M. Arcan, J. P. McCrae, WordNet gloss translation for under-resourced languages using multilingual neural machine translation, in: Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 1–7. URL: https://www.aclweb.org/anthology/W19-7101.

[7] B. R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. S, M. Arcan, M. Zarrouk, J. P. McCrae, Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 56–63. URL: https://www.aclweb.org/anthology/W19-6809.

[8] S. S.Anbukkarasi, Analyzing sentiment in tamil tweets using deep neural network, IEEE Xplore (2020).

[9] S. K. P. Shriya Seshadri, Anand Kumar Madasamy, Analyzing sentiment in indian languages micro text using recurrent neural network, IIOABJ (2016).

[10] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, J. P. McCrae, A survey of current datasets for code-switching research, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, 2020, pp. 136–141.

[11] R. Priyadharshini, B. R. Chakravarthi, M. Vegupatti, J. P. McCrae, Named entity recognition for code-mixed indian corpus using meta embedding, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, 2020, pp. 68–72.

[12] D. Vijay, A. Bohra, V. Singh, M. Akhtar, Syed S.and Shrivastava, Corpus creation and emotion prediction for hindi-english code-mixed social media text, in: NAACL-HLT 2018, NAACL, 2018, pp. 128–135.

[13] P. Ranjan, B. Raja, R. Priyadharshini, R. C. Balabantaray, A comparative study on code-mixed data of Indian social media vs formal text, in: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 2016, pp. 608–611. doi:10.1109/IC3I.2016.7918035.

[14] R. Padmamala, V. Prema, Sentiment analysis of online tamil contents using recursive neural network models approach for tamil language, in: 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), IEEE, 2018, pp. 28–31.

[15] A. S, Github Code, 2020 (accessed October 15, 2020). URL: https://github.com/AnbukkarasiS/Dravadian-Codemix.