JUNLP@Dravidian-CodeMix-FIRE2020: Sentiment Classification of Code-Mixed Tweets using Bi-Directional RNN and Language Tags

Sainik Kumar Mahata, Dipankar Das and Sivaji Bandyopadhyay

Computer Science and Engineering, Jadavpur University, Kolkata, India

Abstract

Sentiment analysis has been an active area of research in the past two decades and recently, with the advent of social media, there has been an increasing demand for sentiment analysis on social media texts. Since the social media texts are not in one language and are largely code-mixed in nature, the traditional sentiment classification models fail to produce acceptable results. This paper tries to solve this very research problem and uses bi-directional LSTMs along with language tagging, to facilitate sentiment tagging of code-mixed Tamil texts that have been extracted from social media. The presented algorithm, when evaluated on the test data, garnered precision, recall, and F1 scores of 0.59, 0.66, and 0.58 respectively.

Keywords

Sentiment Analysis, LSTM, Language Tagging

1. Introduction

Sentiment analysis is the interpretation and classification of emotions (positive, negative, and neutral) within text data using text analysis techniques. It is one of the most important research areas in the domain of Natural Language Processing (NLP) and has garnered much attention in the recent past. However, with the advent of social media, research has become even more wide-spread [1, 2] as it takes into account conversations of customers around the social space and puts them into context. But, in the context of the Indian subcontinent, much research has been focused on sentiment classification of social media texts that are generally code-mixed in nature. This is because India has a linguistically diverse diaspora and a large number of the Indian population is comfortable in more than one language [3]. This leads to communication in sentences, which contain more than one language in the same phrase [4]. Furthermore, words of different languages are generally written in Roman Script, which leads to the formation of a complex syntax structure which is difficult to parse with traditional NLP tools.

This paper aims to solve this research problem and uses Bi-Directional LSTMs [5] to tag the texts with its respective sentiment. Language tagging of individual words was used as additional features while training our classification model. Moreover, the training corpus was passed through FastText [6] embedding, to map the semantically similar words in a common

FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India
EMAIL: sainik.mahata@gmail.com (S. K. Mahata); dipankar.dipnil2005@gmail.com (D. Das); sivaji.cse.ju@gmail.com (S. Bandyopadhyay)

3D space. The designed model was evaluated on the test data and garnered an F1 score of 0.58. The code of our developed model is available as a git repository here.

This model was designed as a part of the "Dravidian-CodeMix - FIRE 2020[1]" shared task and was evaluated for English-Tamil code-mixed texts. The goal of this task was to identify sentiment polarity of the code-mixed dataset of comments/posts in Dravidian Languages (Malayalam-English and Tamil-English) collected from social media.

The rest of the paper is organized as follows. Section 2 deals with the description of the training data that was used to build the proposed sentiment classification system. Section 3 describes the proposed model and the various sub-models used to build the system and will be followed by the evaluation of the model in Section 4.

2. Data

The organizers provided us with Tamil-English and Malayalam-English code-mixed text data, derived from YouTube video comments. The dataset contained all the three types of code-mixed sentences – Inter-Sentential switch, Intra-Sentential switch, and Tag switching and had 5 output labels; Positive, Negative, Mixed Feelings, Not Tamil, and Unknown State. Most comments were written in Roman script with either Tamil / Malayalam grammar with English lexicon or English grammar with Tamil / Malayalam lexicon. Some comments were written in Tamil / Malayalam script with English expressions in between. We participated in the sentiment classification of the code-mixed English-Tamil text task only. The English-Tamil dataset was divided into training, validation and test data which had 11,335, 1,260 and 3,149 code-mixed sentence instances respectively..

3. Framework

Initially, the training and validation data, without the sentiment labels, were merged together, tokenized using the NLTK[2] library, and this was used to extract word vectors of size 100, using the FastText[3] embedding. The skip-gram model was used instead of the continuous-bag-of-words (CBOW) model as skip-gram works best for low data sizes. The model took into account character n-grams from 3 to 6 characters.

After this step, the data was preprocessed to find out the language tags of individual words in a sentence. For this, the nltk corpus was used. The words were given as input and were tagged as English or non-English by the algorithm.

Thereafter, vectors of sentences of the train and validation dataset were extracted from the trained embedding. The language tags and the words vectors were merged together using a Concatenation layer and were given as input to a Bi-Directional LSTM cell. The context vector was then mapped to the output labels with the help of a Dense layer.

The schematic of the model is shown in Figure 1. Other parameters of the model are as follows.

[1] https://dravidian-codemix.github.io/2020/

[2] https://www.nltk.org/

[3] https://fasttext.cc/

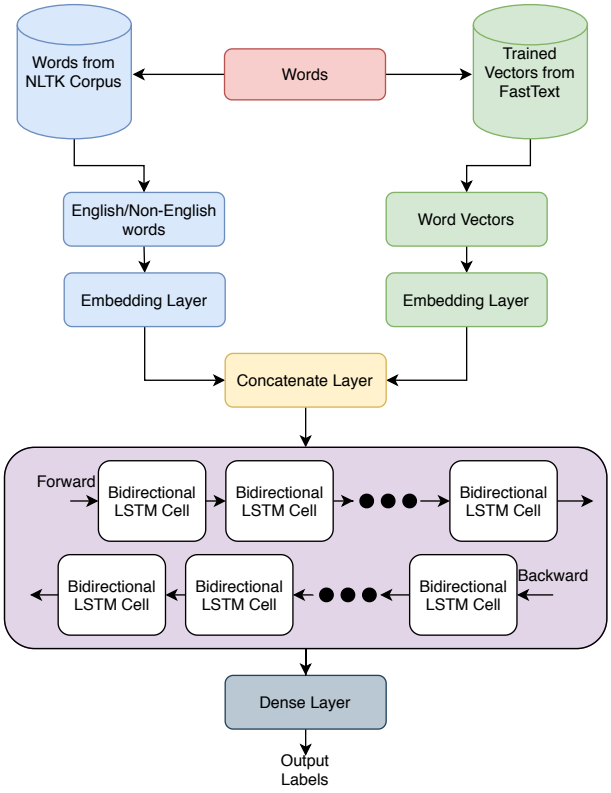

Figure 1: Code-Mixed Sentiment Analysis model.

- batch size: 32
- epochs: 50
- optimizer: adam
- loss: sparse categorical cross-entropy
- validation split: 0.1

4. Evaluation

On testing the model on the initial test dataset, the model garnered accuracy and F1-Score of 70.42% and 0.63 respectively. We also trained three other models, where the basic architecture was the same, the difference being the usage of LSTM/Bi-Directional LSTM and language tag features. The models were

- Bidirectional LSTM without the language tag feature.
- LSTM with the language tag feature.
- LSTM without the language tag feature.

The accuracy and F1-Score of every model are shown in Table 1.

Table 1
Comparison of accuracy scores of all the developed models.

Model	Bi-LSTM+ln tag	Bi-LSTM	LSTM+ln tag	LSTM
Accuracy	70.42%	70.82%	70.62%	70.22%
F1-Score	0.63	0.61	0.62	0.62
Precision	0.62	0.59	0.63	0.62
Recall	0.70	0.71	0.71	0.70

Table 2
Final evaluation result by the Organizers.

Team Name	Precision	Recall	F1-Score
JUNLP	0.59	0.66	0.58

We chose the Bi-Directional LSTM model with language tag features as the final model, as it has the best F1-Score out of all the developed models.

The system was submitted to the organizers and was evaluated using the gold standard dataset, developed by the organizers. The results of the evaluation are shown in Table 2.

5. Conclusion

In the current work, we attempted to solve the problem of Sentiment Analysis of code-mixed English-Tamil data, while participating in the Dravidian-Code-Mixed-FIRE2020 shared task. Our system was based on using Bi-Directional LSTM along with Language Tag features. Also, FastText embedding was used to generate word vectors to train the model. Our system, when evaluated by the organizers garnered an F1 score of 0.58. As future work, we would like to increase this data, use state-of-the-art Neural Network architectures, like BERT, RoBERTa, etc., on this data, taking into advantage the concept of matrix and embedded language, SentiWordNet, and other NLP features.

6. Acknowledgment

This work is supported by Digital India Corporation, MeitY, Government of India, under the Visvesvaraya PhD Scheme for Electronics & IT.

References

[1] S. K. Mahata, S. Makhija, A. Agnihotri, D. Das, Analyzing code-switching rules for english–hindi code-mixed text, in: J. K. Mandal, D. Bhattacharya (Eds.), Emerging Technology in Modelling and Graphics, Springer Singapore, Singapore, 2020, pp. 137–145.

[2] A. Garain, S. K. Mahata, D. Das, JUNLP at SemEval-2020 task 9:Sentiment Analysis of Hindi-English code mixed data using Grid Search Cross Validation, in: Proceedings of

the 14th International Workshop on Semantic Evaluation (SemEval-2020), Association for Computational Linguistics, Barcelona, Spain, 2020.

[3] S. K. Mahata, S. Mandal, D. Das, S. Bandyopadhyay, Code-mixed to monolingual translation framework, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 30–35. URL: https://doi.org/10.1145/3368567.3368579. doi:10.1145/3368567.3368579.

[4] S. K. M. Soumil Mandal, D. Das, Preparing bengali-english code-mixed corpus for sentiment analysis of indian languages, in: K. Shirai (Ed.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France, 2018.

[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735.

[6] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).

[7] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

[8] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[9] 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://www.aclweb.org/anthology/2020.sltu-1.28.

[10] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, 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. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

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