A Sequential DNN for Sentiment Analysis of Dravidian Code-Mixed Language Comments on YouTube

Aaron Samuel. A[1,*,†], Lavanya Sambath Kumar[1,†], Subalalitha Chinnaudayar Navaneethakrishnan[1,†] and Ratnasingam Sakuntharaj[2,†]

[1]SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Tamil Nadu, India
[2]Eastern University, Sri Lanka

Abstract

A method for determining if a block of text is positive, neutral, or negative is sentiment analysis. As code-mixed material in many native languages is becoming increasingly widespread, there is also an increasing need for intense research in order to produce satisfactory results. This research paper aims to classify the sentiments from a data set of comments/posts into pre-defined classes belonging to the code-mixed text in Tamil, Malayalam, and Kannada by utilizing the Sequential Deep Learning model on the code-mixed data set. The sequential model achieved an f1-score of 0.20 for Tamil-English, 0.48 for Malayalam-English, and 0.47 for Kannada-English data sets. The results were submitted to the competition 'Shared Task on Sentiment Analysis and Homophobia detection of YouTube comments in Code-Mixed Dravidian Languages' organized by DravidianLangTech.

Keywords

Sentiment Analysis, Sequential model, Deep Neural Network

1. Introduction

Sentiment analysis at the word level examines how the words or phrases in the text are oriented and how that affects the overall mood, whereas sentiment analysis at the sentence level examines sentences that represent one perspective and makes an effort to identify its direction. The foundation of a lexicon-based method is a corpus or list of words with a particular difference. Then, an algorithm searches for specific words, measures their weight or counts them, and determines the overall duality of the text [2, 3, 4].

In a country where several different languages are spoken, code-mixing becomes commonplace. People who live in multilingual countries employ code-mixed discourse when interacting online and in person. Code-mixing is "the incorporation of linguistic forms from one language, such as phrases, syllables, and morphemes into an expression of a different language." There have already been many experiments that were conducted to make use of sentiment analysis in monolingual texts and they have been successful [15, 25]. But there have been far fewer

FIRE 2022 - Forum for Information Retrieval Evaluation, December 9-13, 2022, Kolkata, India.
*Corresponding author.
†These authors contributed equally.
lavanyas6@srmist.edu.in (L. S. Kumar)

studies conducted for the sentiment analysis of code-mixed languages [16]. The task given in the competition is to categorize the feelings expressed in the code-mixed language data set's YouTube comments. The objective of our study is to classify YouTube comments into the following classes: Positive, negative, unknown_state, mixed-feelings, or if the word is not in that respective language of that particular code-mixed language.
Social media corporations have always been required to fund/contribute to sentimental analysis research to protect social media users from cyberbullying. There have been numerous studies that have looked into sentiment analysis models. However, just a few research papers have looked into the use of Emoji characters on social media [1]. Emojis can significantly change a message when used out of context [17, 20]. The number of studies done on sentiment analysis of code-mixed formats has increased recently [5, 6, 7, 9, 21, 22, 23, 24]. Machine learning models are known as "sequence models" input or output data in a sequential fashion. Time-series data, snippets of audio and video clips, text streams, and other types of data are all examples of sequential data. In sequence models, recurrent neural networks (RNNs) are a widely used technique [11]. Research into distinct sequential data, including time-series data, text expressions, and also other sequential data led to the development of sequence models. While these models perform better with sequential data, CNN (Convolutional Neural Network) models are better suited to handle spatial data. [18]. In the current study, Dravidian languages have been code-mixed with English, such as "Tamil", "Malayalam" and "Kannada" [12, 13, 14]. The data set that was utilized for this study is a component of Task A from the task "Sentiment Analysis and Homophobia Detection of YouTube Comments". The current paper categorizes each YouTube remark at the message level into one of the following categories: "Positive," "Negative," "Not Tamil/Malayalam/Kannada," "Unknown_State," and "Mixed Feelings."

2. Data set Description

The data set used here is provided by Task A proposed [19]. It is a collection of YouTube comments in Tamil, Malayalam, and Kannada (data for all 3 from [8, 10]. The comments are all heavily code-mixed and that data has been classified into five classes which are, unknown_state, Mixed feelings, positive, negative, and not-(Tamil or Malayalam, or Kannada). The description and split of the data set are illustrated in Table 1.

Table 1
Description of the train, validation, test data sets, and all the labels of Tamil, Malayalam and Kannada

Language	Train Data set	Validation Data set	Test Data set
Tamil	35656	3962	649
Malayalam	15888	1766	1962
Kannada	6212	691	768

3. Text Pre-processing

Due to the code-mixing and blatant disregard for grammatical rules in the data set downloaded from YouTube. To use the data set effectively, the following processes are applied.

- Initially emojis, special characters, numbers, and punctuation were all eliminated as they have no functional use to a statement.
- We lowercase all the characters and replace usernames with empty characters.
- We then split the comment into tokens.
- Next, we made a flat list of all words from the corpus and then we computed the number of occurrences of all the words in the corpus.
- Padding is applied to the corpus next.
- We created a sequential model and passed the required layers to it.
- To avoid over-fitting our model, we have used Early Stopping as well.

4. Proposed Methodology

A Sequential DNN was developed for the sentiment analysis tasks. These networks received their input from the embedding vectors. The text indexes are transformed into dense vectors with defined sizes. The input length, embedding initializer, and embedding Regularizer assigned were all of ”maximum length,” ”orthogonal,” and ”L2 Regularizer,” respectively. Next, we added the LSTM layer and wrapped the layer with Bidirectional. The bi-directionality of a Keras layer was added to the model by implementing tf.keras.layers.bidirectional to the model. Finally, we used the Dense layer to classify the data into the 5 classes and used the ‘softmax’ activation function. We compiled our model and defined the loss function, optimizer, and metrics. We pick “Categorical Cross-Entropy” as the loss function because the provided task requires multi-class categorization. For the given task, we used the default optimizer “Adam” and set the learning rate to 0.01.

Table 2
Description of the various parameters for the Sequential model.

Parameter	Value
Number of LSTM units	32
Activation Function	Softmax
Embedding Initializer	Orthogonal
Embedding Regularizer	L2
Batch-size	256
Optimiser	Adam
Learning Rate	0.01
Loss Function	Categorical Cross-entropy
Epochs	2

The model is trained in order to fine-tune the parameters to produce the desired outputs for a particular input. This is accomplished by putting inputs into the input layer, receiving an output, computing the loss function using the output, and then fine-tuning the model parameters using back-propagation. As a result, the model's parameters will be fit and matched to the data. The batch size for the model while fitting was 256 and the number of epochs was 2. Table 2 displays the several parameters used in the Sequential model.

5. Implementation

All the required modules and packages like TensorFlow, pandas, NumPy, Regular Expression, Natural Language Toolkit, scikit-learn, etc. are all imported to the notebook file. The feature extraction and model training is done in Python using the scikit-learn library. The text data is transformed into TF-IDF feature vectors using the scikit-learn Tfidf Vectorizer.

6. Results

The metrics precision, recall, and f1-score have been calculated for the code-mixed data sets. A precision of 0.22, a recall of 0.18, and an f1-score of 0.20 were recorded for the Tamil-English data set as shown in Table 3.

Table 3
Description of the comparison of accuracy metrics between our team and other teams for Tamil-English

Team Name	Precision	Recall	F1-score
SRMNLP	0.340	0.330	0.270
BharathNLP	0.190	0.220	0.190
bilstm	0.220	0.190	0.190
SSN-CSE	0.220	0.260	0.170
Sentiment	0.240	0.220	0.170
MUCS	0.240	0.190	0.160
Fnet	0.150	0.130	0.130
JPMCAI	0.020	0.160	0.020
Task Masters (Our Team)	0.220	0.180	0.200

A precision of 0.51, a recall of 0.57, and an f1-score of 0.48 were recorded for the Malayalam-English data set as shown in Table 4.

A precision of 0.48, a recall of 0.50, and an f1-score of 0.47 were recorded for the Kannada-English data set as shown in Table 5. The comparison between our results and other competitors' results was shown in Tables 3,4 and 5 respectively. As a result, we tested with the sequential DNN model for the three code-mixed data sets. Any language can be used with this approach because it is language-independent.

Table 4
Description of the comparison of accuracy metrics between our team and other teams for the Malayalam-English

Team Name	Precision	Recall	F1-score
IRLAB	0.670	0.670	0.660
Fnet	0.660	0.620	0.640
Sentiment	0.620	0.630	0.630
MUCS	0.610	0.610	0.610
NITK	0.600	0.600	0.600
SRMNLP	0.610	0.550	0.570
lone_warrior	0.520	0.590	0.520
Bilstm	0.490	0.580	0.500
BharathNLP	0.160	0.270	0.200
JPMCAI	0.340	0.200	0.140
SSN-CSE	0.090	0.140	0.110
Task Masters (Our Team)	0.510	0.570	0.480

Table 5
Description of the comparison of accuracy metrics between our team and other teams for the Kannada-English

Team Name	Precision	Recall	F1-score
IRLAB	0.560	0.560	0.550
Sentiment	0.520	0.500	0.510
lone_warrior	0.470	0.510	0.480
NITK	0.480	0.500	0.480
Fnet	0.500	0.490	0.480
AI Defenders	0.490	0.480	0.480
SRMNLP	0.540	0.440	0.460
JPMCAI	0.550	0.430	0.450
MUCS	0.470	0.460	0.440
Bilstm	0.480	0.500	0.430
QWERTY	0.460	0.350	0.350
BharataNLP	0.290	0.330	0.300
SSN-CSE	0.120	0.170	0.110
Task Masters (Our Team)	0.480	0.500	0.470

7. Conclusion

Comparatively, we can see that the Malayalam-English data set had the highest f1-score, precision, and recall, and the Tamil-English data set had the lowest f1-score, precision, and recall with Kannada-English data set having precision and f1-score just slightly lesser than that of Malayalam. However, when compared to the other two data sets, the Tamil-English data set appeared to perform poorly with the model. However, it should be noticed that compared

to the other classes, the 'Positive' class in the Tamil-English data set has significantly too many instances. The imbalance in the data has resulted in lower accuracy. The training and development data was substantial in comparison to the other two languages, which contributed to the poor scores compared to Malayalam-English and Kannada-English data sets.

References

[1] Shiha, M., & Ayvaz, S. (2017). The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng.(IJCEE.), 9(1), 360-369.

[2] Melville, P., Gryc, W., & Lawrence, R. D. (2009, June). Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1275-1284).

[3] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011, August). User-level sentiment analysis incorporating social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1397-1405).

[4] Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240).

[5] Prabhu, A., Joshi, A., Shrivastava, M., & Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. arXiv preprint arXiv:1611.00472.

[6] Kumar, A., Saumya, S., & Singh, J. P. (2020). NITP-AI-NLP@ Dravidian-CodeMix-FIRE2020: A Hybrid CNN and Bi-LSTM Network for Sentiment Analysis of Dravidian Code-Mixed Social Media Posts. In FIRE (Working Notes) (pp. 582-590).

[7] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210.

[8] Priyadharshini, R., Chakravarthi, B. R., Thavareesan, S., Chinnappa, D., Thenmozhi, D., & Ponnusamy, R. (2021, December). Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada. In Forum for Information Retrieval Evaluation (pp. 4-6).

[9] Suryawanshi, S., & Chakravarthi, B. R. (2021, April). Findings of the shared task on Troll Meme Classification in Tamil. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages (pp. 126-132).

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

[11] Denoyer, L., & Gallinari, P. (2014). Deep sequential neural network. arXiv preprint arXiv:1410.0510.

[12] Chakravarthi, B. R., Arcan, M., & McCrae, J. P. (2019). 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.

[13] Andrew, J. (2020, December). JudithJeyafreeda@ Dravidian-CodeMix-FIRE2020:: Senti-

ment Analysis of YouTube Comments for Dravidian Languages. In Forum for Information Retrieval Evaluation.

[14] Kumar, A., Saumya, S., & Singh, J. P. (2020). NITP-AI-NLP@ Dravidian-CodeMix-FIRE2020: A Hybrid CNN and Bi-LSTM Network for Sentiment Analysis of Dravidian Code-Mixed Social Media Posts. In FIRE (Working Notes) (pp. 582-590).

[15] Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2015, September). Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 2-8).

[16] Patra, B. G., Das, D., & Das, A. (2018). Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. arXiv preprint arXiv:1803.06745.

[17] Guibon, G., Ochs, M., & Bellot, P. (2016, June). From emojis to sentiment analysis. In WACAI 2016.

[18] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9(4), 611-629.

[19] Chakravarthi, B.R., Priyadharshini, R., Ponnusamy, R., Kumaresan, P.K., Sampath, K., Thenmozhi, D., Thangasamy, S., Nallathambi, R. and McCrae, J.P., 2021. Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments. arXiv preprint arXiv:2109.00227.

[20] Shumugavadivel, Kogilavani and Subramanian, Malliga and Kumaresan, Prasanna Kumar and Chakravarthi, Bharathi Raja and B, Bharathi and Chinnaudayar Navaneethakrishnan, Subalalitha and S.K, Lavanya and Mandl, Thomas and Ponnusamy, Rahul and Palanikumar, Vasanth and Balaji J, Manoj. Overview of the Shared Task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages. In proceedings of dravidiancodemix-2022.

[21] Shanmugavadivel, K., Sampath, S.H., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P.K. and Priyadharshini, R., 2022. An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. Computer Speech Language, p.101407.

[22] Chakravarthi, B.R., Hande, A., Ponnusamy, R., Kumaresan, P.K. and Priyadharshini, R., 2022. How can we detect Homophobia and Transphobia? Experiments in a multilingual code-mixed setting for social media governance. International Journal of Information Management Data Insights, 2(2), p.100119.

[23] Chakravarthi, B.R., 2022. Hope speech detection in YouTube comments. Social Network Analysis and Mining, 12(1), pp.1-19.

[24] Chakravarthi, B.R., 2022. Multilingual hope speech detection in English and Dravidian languages. International Journal of Data Science and Analytics, 14(4), pp.389-406.

[25] Subramanian, M., Ponnusamy, R., Benhur, S., Shanmugavadivel, K., Ganesan, A., Ravi, D., Shanmugasundaram, G.K., Priyadharshini, R. and Chakravarthi, B.R., 2022. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. Computer Speech Language, 76, p.101404.