An ensemble-based model for sentiment analysis of Dravidian code-mixed social media posts

Abhinav Kumar[1], Sunil Saumya[2] and Jyoti Prakash Singh[3]

[1]Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India
[2]Department: Computer Science & Engineering, Indian Institute of Information Technology Dharwad, India
[3]Department: Computer Science & Engineering, National Institute of Technology Patna, India

Abstract

Sentiment analysis is highly important in social media monitoring since it helps us to see how the general population feels about a certain issue. Several studies have been published in recent years that attempt to extract sentiment from social media messages. However, the majority of the work is verified using just English language datasets. Machine learning algorithms do not perform equally well when social media posts are written in multilingual and code-mixed script. This paper presents an ensemble-based model to classify Kannada-English, Malayalam-English, and Tamil-English social media postings into five different sentiment classes using character-level TF-IDF features as input. The proposed ensemble-based model achieved the weighted F_1-scores of 0.62, 0.73, and 0.62 for Kannada-English, Malayalam-English, and Tamil-English datasets, respectively. The code for the proposed models is available at: https://github.com/Abhinavkmr/Dravidian-Sentiment-Analysis-.git

Keywords

Sentiment analysis, Code-mixed, Kannada-English, Tamil-English, Malayalam-English, YouTube, Machine learning, Deep learning

1. Introduction

Sentiment analysis helps in the recognition of opinions or responses on a given topic. Due to its enormous influence on companies such as e-commerce, recommendation systems, hate speech detection [1, 2], and disaster management [3, 4], and social media monitoring, it is one of the most explored subjects in natural language processing. English is the most popular and widely accepted language on the world, and it is widely used over Internet. However, in a nation like India, where over 400 million people use the internet, people utilise more than one language to express themselves, resulting in a new code-mixed language [5]. Dravidian languages such as Malayalam and Kannada are spoken in the Indian states of Kerala and Karnataka. Tamil, which is spoken by Tamil people in India, Singapore, and Sri Lanka, is another well-known Dravidian language in India's southern area. People on social media commonly use Roman script to write these Dravidian languages since it is easy to do so with the keyboards accessible on their devices. The majority of existing models trained to extract sentiment from a single language fail to grasp the semantics of a code-mixed language. Due to its multilingual character, extracting feelings from code mixed user-generated texts becomes more challenging [6, 7].

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

abhinavanand05@gmail.com (A. Kumar); sunil.saumya@iiitdwd.ac.in (S. Saumya); jps@nitp.ac.in (J. P. Singh)

CEUR Workshop Proceedings (CEUR-WS.org)

The sentiment analysis of code-mixed language has recently caught the interest of the research community [8, 9]. Kumar et al. [9] proposed a hybrid CNN and Bi-LSTM Network to classify social media posts into different sentiment classes. Mahata et al. [10] proposed bi-directional LSTM with language tagging to classify Tamil-English and Malayalam-English code-mixed social media posts into different sentiment classes. Sharma and Mandalam [11], on the other hand, employed sub-word level representation to capture text sentiment and implemented an LSTM network to classify Tamil-English and Malayalam-English social media posts into the different polarity classes. Patra et al. [12] presented a model for Bengali-English code mixed data using a support vector machine with character n-grams features. To extract emotions from Hinglish and Spanglish (Spanish + English) data, Advani et al. [13] utilised logistic regression using handcrafted lexical and semantic features. Similarly, On Hinglish data, Goswami et al. [14] presented a morphological attention model for sentiment analysis.

This paper presents an ensemble-based model that uses character-level TF-IDF features to classify Kannada-English, Malayalam-English, and Tamil-English social media posts into five different sentiment classes. The proposed model is validated on the dataset provided in the *DravidianCodeMix FIRE 2021* [15, 16] track. The dataset includes five distinct sentiment classes, including ”positive,” ”negative,” ”mixed feelings,” ”unknown state,” and ”if the post is not in the mentioned Dravidian languages.”

The rest of the sections are organized as follows: the proposed methodology is explained in Section 2. The experimental findings are listed in Section 3 and Section 4 concludes the paper by highlighting the main findings of this study.

2. Methodology

The systematic diagram of the proposed ensemble-based model for the Kannada-English language can be seen in Figure 1 whereas, the proposed model for Malayalam-English and Tamil-English can be seen in Figure 2. The proposed model is validated with the datasets given in the DravidianCodeMix FIRE 2021 competition [16]. The overall data statistic for Kannada-English [17], Malayalam-English [18], and Tamil-English [19] can be seen in Table 1.

Table 1
Overall data statistic for Kannada, Malayalam, and Tamil dataset

Class	Kannada-English			Malayalam-English			Tamil-English		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Mixed-feelings	574	52	65	926	102	134	4,020	438	470
Negative	1,188	139	157	2,105	237	258	4,271	480	477
Positive	2,823	321	374	6,421	706	780	20,070	2,257	2,546
Unknown state	711	69	62	5,279	580	643	5,628	611	665
Not-Kannada	916	110	110	-	-	-	-	-	-
Not-Malayalam	-	-	-	1,157	141	147	-	-	-
Not-Tamil	-	-	-	-	-	-	1,667	176	244
Total	6,212	691	768	15,888	1,766	1,962	35,156	3,962	4,402

Extensive experiments were carried out with a variety of popular machine learning classifiers using various combinations of one-to-six gram word-level and character-level TF-IDF features.

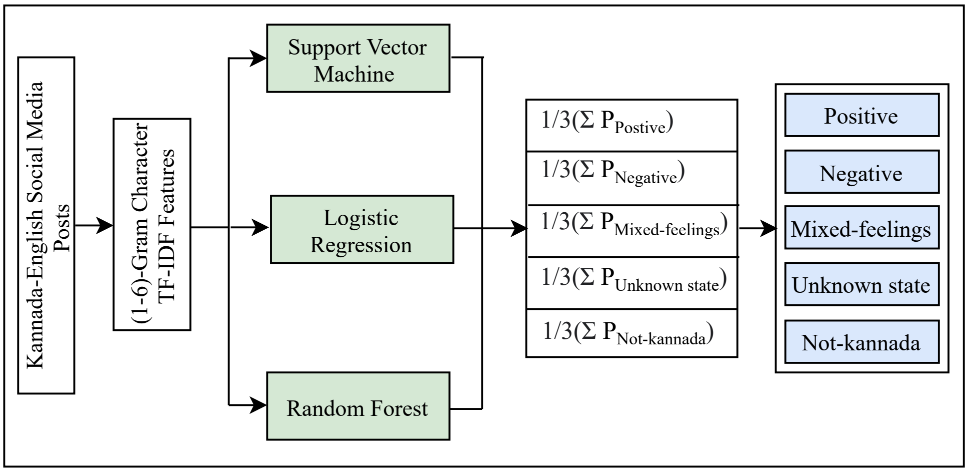


Figure 1: Proposed model for the Kannada-English language

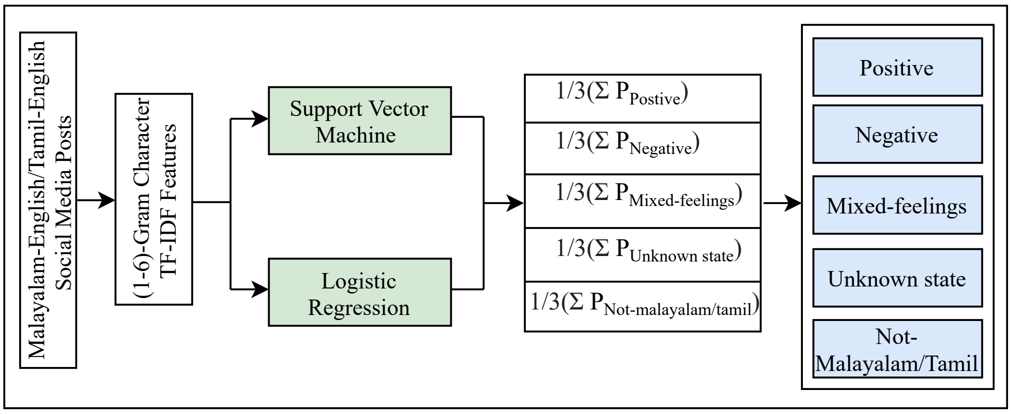


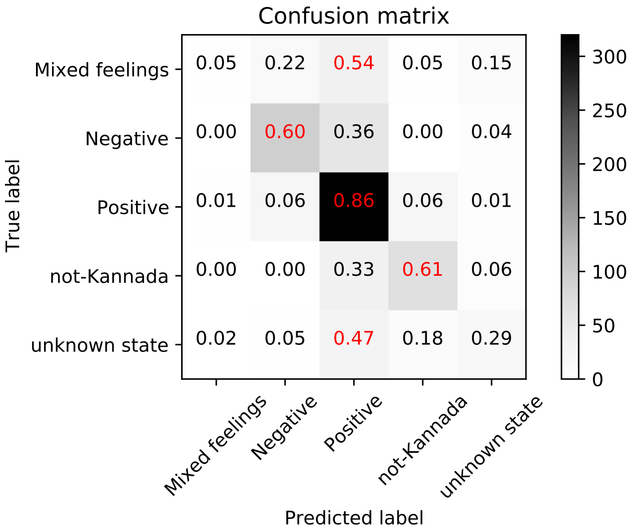
Figure 2: Proposed model for the Malayalam-English and Tamil-English languages

We found that the ensemble of Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) performed best on the Kannada-English dataset, while the ensemble of SVM and LR performed best on the Malayalam-English and Tamil-English datasets. The proposed models are described in detail in the following sections.

- **Kannada-English:** An ensemble-based model is proposed containing SVM, LR, and RF in parallel (see Figure 1). This ensemble-based model uses one to six-gram character TF-IDF features to predict the probability for each of the classes. To choose the suitable character n-gram range, extensive experimentation was performed with one-gram to six-gram character-level TF-IDF features. We found first 50,000 one to six-gram character-level TF-IDF features were performed better than the other combination of character-level n-gram TF-IDF features. The probabilities of all the three classifiers are then averaged class-wise to get the final class probability. The final class for the post is assigned that has the highest average class probability.

Table 2
Performance of the proposed model for Kannada, Malayalam, and Tamil social media posts

Class	Kannada-English			Malayalam-English			Tamil-English		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Mixed-feelings	0.50	0.05	0.08	0.55	0.30	0.39	0.37	0.15	0.21
Negative	0.70	0.60	0.65	0.69	0.57	0.63	0.47	0.33	0.39
Positive	0.67	0.86	0.75	0.76	0.84	0.80	0.70	0.90	0.79
Unknown state	0.39	0.29	0.33	0.71	0.76	0.74	0.50	0.34	0.41
Not-Kannada	0.64	0.61	0.62	-	-	-	-	-	-
Not-Malayalam	-	-	-	0.83	0.74	0.78	-	-	-
Not-Tamil	-	-	-	-	-	-	0.73	0.53	0.61
Weighted Avg.	0.64	0.65	0.62	0.73	0.73	0.73	0.61	0.65	0.62



	Mixed feelings	Negative	Positive	not-Kannada	unknown state
Mixed feelings	0.05	0.22	0.54	0.05	0.15
Negative	0.00	0.60	0.36	0.00	0.04
Positive	0.01	0.06	0.86	0.06	0.01
not-Kannada	0.00	0.00	0.33	0.61	0.06
unknown state	0.02	0.05	0.47	0.18	0.29

Figure 3: Confusion matrix for the Kannada-English dataset

- **Malayalam-English & Tamil-English:** The proposed ensemble-based model contains support vector machine and logistic regression in parallel (see Figure 2). For the Malayalam-English language, first, 30,000 one to six-gram character-level TF-IDF features performed best in comparison to other combinations of n-gram features. For the Tamil-English language, the first 15,000 one to six-gram character-level TF-IDF features performed best in comparison to other combinations of n-gram features. Similar to the previous model (Figure 1) class-wise averaged probabilities were calculated and the final class label is assigned that has the highest average class probability.

3. Results and Analysis

Precision, recall, and the $F1$-score are utilised to assess the suggested ensemble-based model's performance. The confusion matrix and AUC-ROC curve are also presented to highlight the

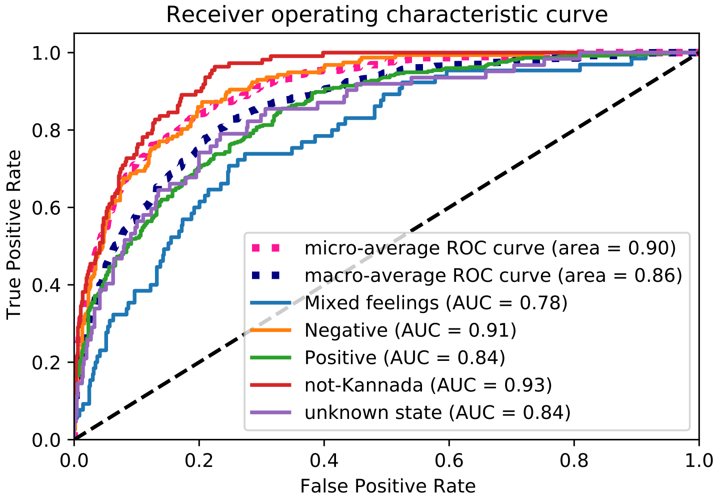


Figure 4: ROC curve for the Kannada-English dataset

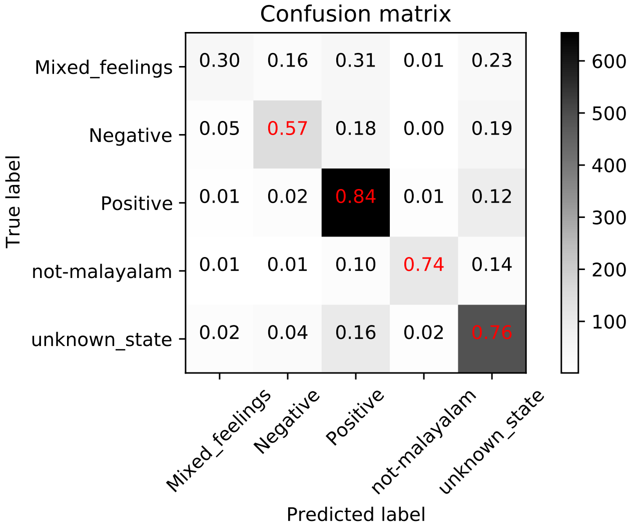


Figure 5: Confusion matrix for the Malayalam-English dataset

model's performance in addition to these measures. Table 2 shows the outcomes of the suggested model for the Kannada-English, Malayalam-English, and Tamil-English languages.

The suggested ensemble-based model has a weighted precision of 0.64, recall of 0.65, and $F1$-score of 0.62 for the Kannada-English dataset. Figures 3 and 4 show the ROC curve and confusion matrix for the Kannada-English dataset, respectively. The suggested ensemble-based model had a weighted precision, recall, and $F1$-score of 0.73 for the Malayalam-English dataset. Figures 5 and 6 show the confusion matrix and ROC curve for the Malayalam-English dataset, respectively. The suggested ensemble-based model achieved a weighted precision of 0.61, recall

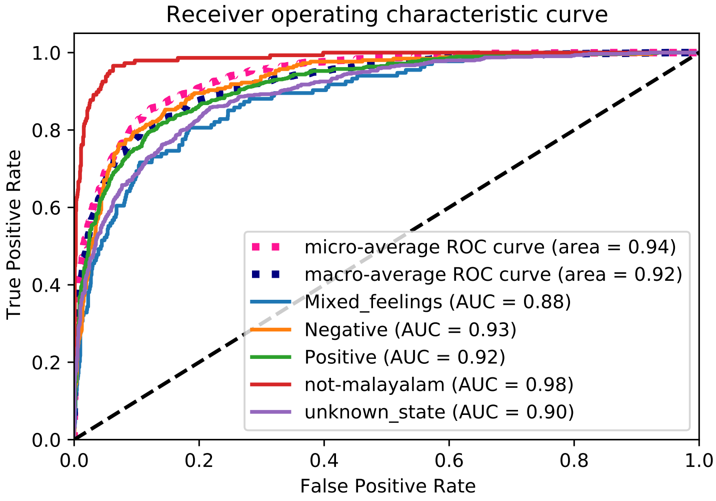


Figure 6: ROC curve for the Malayalam-English dataset

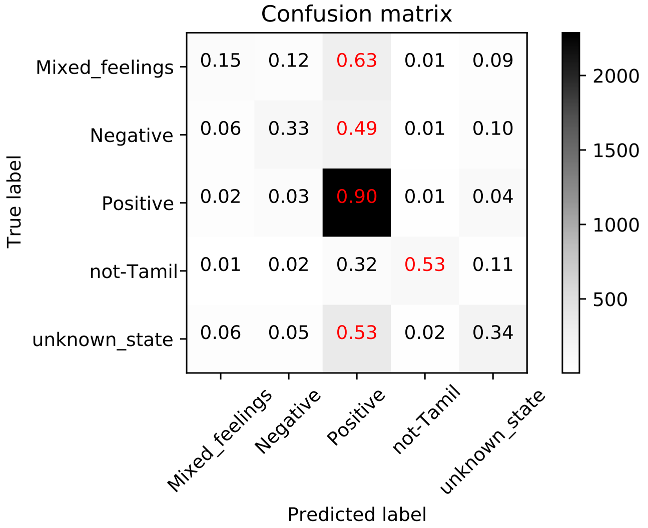


Figure 7: Confusion matrix for the Tamil-English dataset

of 0.65, and $F1$-score of 0.62, respectively, on the Tamil-English dataset. Figures 7 and 8 show the confusion matrix and ROC curve for the Tamil-English dataset, respectively.

4. Conclusion

Sentiment analysis of social media messages is an essential task in natural language processing, which analyses social discussions and feedback to discover the deeper context as they

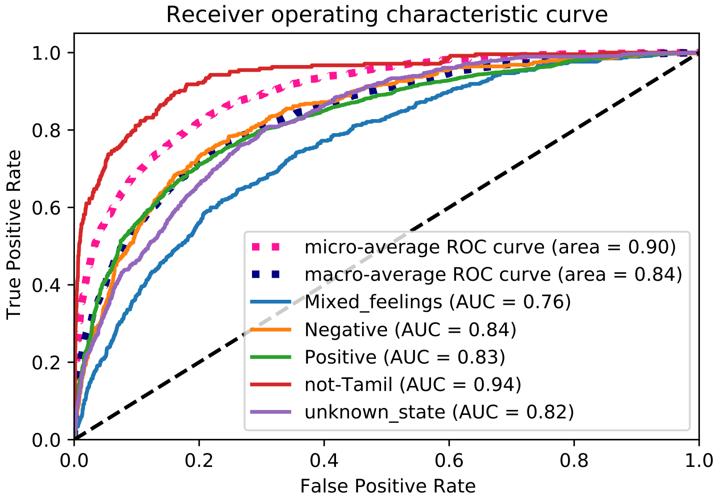


Figure 8: ROC curve for the Tamil-English dataset

pertain to a topic, brand, or theme. This work proposes an ensemble-based model to classify Kannada-English, Malayalam-English, and Tamil-English social media postings into five different sentiment classes. The use of one to six-gram character-level feature performed best with the other combinations of n-gram character-level features. For the Kannada-English, Malayalam-English, and Tamil-English datasets, the suggested ensemble-based model achieved weighted $F1$-scores of 0.62, 0.73, and 0.62, respectively. To improve performance, a robust deep ensemble-based model can be developed in the future by integrating character-level and word-level features.

References

[1] A. K. Mishra, S. Saumya, A. Kumar, IIIT_DWD@ HASOC 2020: Identifying offensive content in indo-european languages (2020).

[2] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in Dravidian code mixed social media text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 36–45.

[3] A. Kumar, J. P. Singh, Y. K. Dwivedi, N. P. Rana, A deep multi-modal neural network for informative twitter content classification during emergencies, Annals of Operations Research (2020) 1–32.

[4] A. Kumar, J. P. Singh, Location reference identification from tweets during emergencies: A deep learning approach, International journal of disaster risk reduction 33 (2019) 365–375.

[5] 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), 2020.

[6] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly,

J. P. McCrae, Dravidiancodemix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text, arXiv preprint arXiv:2106.09460 (2021).

[7] A. Hande, S. U. Hegde, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, S. Thavareesan, B. R. Chakravarthi, Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced Dravidian languages, arXiv preprint arXiv:2108.03867 (2021).

[8] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Davidian Languages in Code-Mixed Text, in: Working Notes of the FIRE 2020. CEUR Workshop Proceedings., 2020.

[9] A. Kumar, S. Saumya, J. P. Singh, 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), 2020, pp. 582–590.

[10] S. Mahata, D. Das, S. Bandyopadhyay, Sentiment classification of code-mixed tweets using bi-directional rnn and language tags, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 28–35.

[11] Y. Sharma, A. V. Mandalam, Bits2020@ Dravidian-CodeMix-FIRE2020: Sub-word level sentiment analysis of Dravidian code mixed data., in: FIRE (Working Notes), 2020, pp. 503–509.

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

[13] L. Advani, C. Lu, S. Maharjan, C1 at SemEval-2020 Task 9: SentiMix: Sentiment analysis for code-mixed social media text using feature engineering, arXiv preprint arXiv:2008.13549 (2020).

[14] K. Goswami, P. Rani, B. R. Chakravarthi, T. Fransen, J. P. McCrae, ULD@ NUIG at SemEval-2020 Task 9: Generative morphemes with an attention model for sentiment analysis in code-mixed text, arXiv preprint arXiv:2008.01545 (2020).

[15] 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.

[16] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, D. Thenmozi, 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.

[17] A. Hande, R. Priyadharshini, B. R. Chakravarthi, KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 54–63. URL: https://www.aclweb.org/anthology/2020.peoples-1.6.

[18] 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.

[19] 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.