# PITS@Dravidian-CodeMix-FIRE2020: Traditional Approach to Noisy Code-Mixed Sentiment Analysis

Nikita Kanwar[a], Megha Agarwal[a] and Rajesh Kumar Mundotiya[b]

[a]*Pratap Institute of Technology & Science, India*
[b]*Indian Institute of Technology (BHU), India*

**Abstract**

Sentiment Analysis (SA) is a process for characterizing the response or opinion by which sentiment polarity of the text is decided. Nowadays, social media is a common platform to convey opinions, suggestions and much more in a user's native language or multilingual in Roman script (for ease). In this task, Malayalam-English and Tamil-English code mixed dataset in the Roman script has provided for SA. To solve this task, we have generated syntax-based features and used trained logistic regression with as an under-sampling technique. We have obtained best $F_1$-score of $0.71$ and $0.62$ on the blind test set of Malayalam-English and Tamil-English code mixed datasets, respectively. The code is available at Github[1].

**Keywords**

Malayalam-English code-mixed, Tamil-English code-mixed, Sentiment Analysis, Logistic Regression,

## 1. Introduction

Sentiment Analysis (SA) is a process for characterizing the response or opinion, which determines sentiment polarity of the text. In the last few years, social media platforms such as Facebook, Twitter and Youtube have become increasingly large, which in turn produced textual data as people express their feelings and opinions by writing reviews, comments on social media.

A large number of texts on such platforms are available in either user's native language, English or a mixture of both. According to Myers-Scotton (1993), code-mixing is defined as – "The interchangeable use of linguistic units, like morphs, words, and phrases from one language to another language while conversation (both speaking and writing)" [1]. The code-mixing text does not follow the formal grammar or even writing script. The use of writing script entirely depends upon the user. Hence, it can be directly stated that the traditional approaches for SA does not provide an effective solution. However, this problem becomes more complicated when code-mixing text follows multilingualism, which is what most Indian users do nowadays. Consequently, processing of code-mixed text has been gaining a propagation of attention and interest in the NLP community [2, 3, 4, 5, 6, 7, 8].

[1]https://github.com/Rajesh-NLP/CODE-MIXED-SENTI-FIRE20

*FIRE 2020: Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India*
EMAIL: rathorenik2@gmail.com (N. Kanwar); megha.agarwal259@gmail.com (M. Agarwal); rajeshkm.mundotiya@gmail.com (R. K. Mundotiya)

In this paper, we have worked with such kinds of code-mixed texts of the Dravidian languages (Malayalam and Tamil) with English for SA by using the traditional approach with syntactic features.

## 2. Related Work

Over the past few years, social media has gained a boost in the code-mixed or code-switched text after getting multilingual support. Therefore, code-mixed SA also has a significant problem. This problem has been solved by two different methods, namely lexicon and machine learning [9]. Sharma et al. [10], Pravalika et al. [11], Baccianella et al. [12] have used lexicon-based approaches for SA on code-mixed dataset. Whereas, traditional machine learning-based approaches such as Naive Bayes, Support Vector Machine, Decision Tree and many more with hand-crafted (Sarkar [13], Baccianella et al. [12]) and syntactic-based features (Chakravarthi et al. [7, 8], Remmiya Devi et al. [14], Kouloumpis et al. [15]) provided significant results on the code-mixed dataset.

However, nowadays, many researchers have been approaching this problem through deep learning methods, which are also able to capture computational aspects to some extent [9]. Mishra et al. [16] has built a multi-layer perceptron and bidirectional-long short term memory (LSTM) with Glove embeddings to perform SA on Hindi-English and Bengali-English datasets. Joshi et al. [17] tried to leverage the subword information in the deep learning-based model on the Hindi-English dataset, which was later extended by Mukherjee [18]. In this extension, the author used an LSTM followed by a Convolutional Neural Network (CNN) for performing joint learning between word and character-level features. These distributional representations capture semantic information at particular extent, i.e. till window size in word embeddings or a certain length of the input sentence by the variants of recurrent neural network due to gradient vanishing problem. Nowadays, the contextual word embeeding techniques are prominent in this case. BERT and ELMo are contextual word embeddings that are introducing new baseline goals for SA. However, the performance of these embeddings with deep-learning models suffers for code-mixed datasets [7, 8].

## 3. Features and Technique

We have generated $40,000$ syntactic features by the combination of word and character-based information. The n-gram technique helps to leverage such information. On each word of the text, unigram and bigram were generated without removing any stop words or stemming. Similarly, n-gram was also implemented at the character level, and the word boundary was kept in mind. These n-gram features were encoded by Term Frequency-Inverse Document Frequency (TF-IDF) to generate feature space, where each YouTube's comment was considered a document.

This feature space was used to train the logistic regression model with an L1-regularizer. However, the provided datasets are imbalanced as shown in Figure 1; hence we used the under-sampling technique, i.e., Tomek's link [19]. Tomek's link calculates the distance among the class-wise samples and finds the nearest samples by using the nearest neighbour supervised technique and removing it from majority classes.

**Table 1**
Statistics of split code-mixed datasets' into Training, Testing and Validation

| Language | Training | Validation | Testing |
|---|---|---|---|
| **Malayalam-English** | 4,851 | 541 | 1,348 |
| **Tamil-English** | 11,335 | 1,260 | 3,149 |

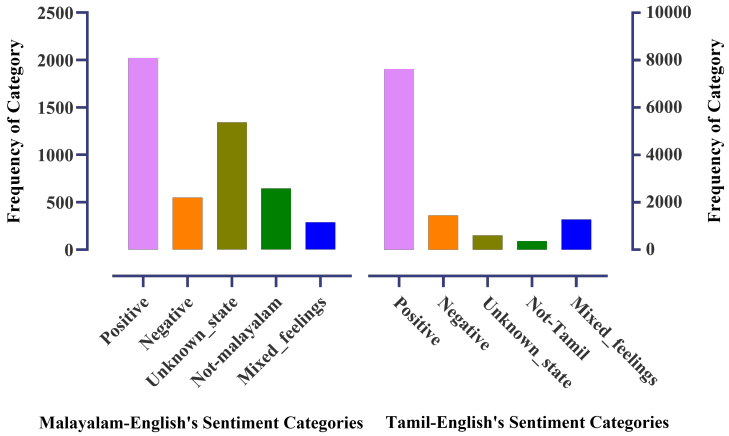


**Figure 1:** Category-wise distribution in Malayalam-English (left) and Tamil-English (right) code-mixed datasets

# 4. Experiment

## 4.1. Dataset

The datasets used in this experiment are collected from YouTube comments in Malayalam-English and Tamil-English as code-mixed in the Roman script, which contains 6,739 and 15,744 texts, respectively. Both the code-mixed datasets follow Tag switching, Intra- and Inter-Sentential switch [7, 8]. The division of datasets to training, validation and testing are summarized in Table 2. These datasets have annotated into five categories, namely Positive, Negative, Mixed-feeling, Unknown-state, Not Malayalam or Not Tamil for the SA. The distribution of categories is imbalanced in the combinations of provided training and validation dataset, as shown in Figure 1.

## 4.2. Settings

Traditional machine learning approaches are a prominent method for providing robust solutions on a scarce dataset. Hence we have performed combinations of features and classification techniques. We have cleaned the dataset by removing emojis and smiles through the tweet-preprocessor[1] before generating the features. Such cleaning degrades the model performance in our experiments. The generated features are word length, character and word n-grams, word

[1]https://pypi.org/project/tweet-preprocessor/

**Table 2**
The obtained **P**recision, **R**ecall and $\mathbf{F}_1$-score on the validation and blind test datasets

| Code-Mixed Dataset | Validation | | | Test | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | $\mathbf{F}_1$ | **P** | **R** | $\mathbf{F}_1$ |
| **Malayalam-English** | 0.69 | 0.70 | 0.69 | 0.70 | 0.71 | 0.71 |
| **Tamil-English** | 0.64 | 0.71 | 0.64 | 0.62 | 0.69 | 0.62 |

repetitions, word count and presence of punctuation used with different classification techniques namely Decision Tree, Support Vector Machine, Catboost, XGBoost, Logistic Regression.

The Logistic Regression with under-sampling technique provides best results on the default value of parameters in the sklearn library[2] on the validation dataset by using the word, and character-based n-gram features. However, the Tamil-English has not shown any effect on model performance. Here, we have considered bigram for word-level and bigram to six-gram for character-level features. Out of the yielded features, $1,000$ word-level and $30,000$ character-level features have been used.

## 5. Results and Analysis

After applying logistic regression on the encoded TF-IDF word and character-based n-gram features, we obtained the $\mathrm{F}_1$-score of $0.69$, $0.71$ and $0.64$, $0.62$ on the validation and blind test dataset of Malayalam-English and Tamil-English, respectively. The evaluations mentioned in Table 2, considers three different metrics, namely Precision, Recall, and $\mathrm{F}_1$-score. From empirical observations of the obtained results, we found that our model correctly classified most of the relevant categories. The category-wise scores on the validation datasets are mentioned in Table 3. From this, we observe that the model faces difficulties while learning for the "Mixed_feeling" category, hence the score of this category is less as compared to other categories. In both datasets, our model has been vastly confused in "Mixed_feelings", and "unknown_state" categories. Most of these categories predicted as "Positive" category in the validation datasets as shown in confusion metrics (in Figure 2), appended in the Appendix section.

## 6. Conclusion

This paper shows that logistic regression with under-sampling technique achieved comparable metric scores on the code mixed sentiment analysis dataset of Malayalam-English and Tamil-English. This technique is also relying on the word and character-level features, hence it provides $0.71$ and $0.62$ as $\mathrm{F}_1$-scores on the blind test set of respective datasets. From empirical observations of the obtained results on the validation set, we found that this technique correctly classifies most of the relevant categories.

[2] https://scikit-learn.org/stable/

**Table 3**
Category wise evaluation scores, obtained on the validation datasets. In Not-Malayalam/Tamil category, Not-Malayalam is used for Malayalam-English datasets, and Not-Tamil is used for Tamil-English

| Dataset | Malayalam-English | | | Tamil-English | | |
|---|---|---|---|---|---|---|
| **Labels** | **P** | **R** | **$F_1$** | **P** | **R** | **$F_1$** |
| **Mixed_feelings** | 0.60 | 0.35 | 0.44 | 0.37 | 0.05 | 0.08 |
| **Negative** | 0.64 | 0.56 | 0.60 | 0.42 | 0.22 | 0.30 |
| **Positive** | 0.76 | 0.78 | 0.77 | 0.74 | 0.96 | 0.83 |
| **Not-Malayalam/Tamil** | 0.73 | 0.76 | 0.74 | 0.83 | 0.54 | 0.66 |
| **Unknown_state** | 0.63 | 0.67 | 0.65 | 0.33 | 0.09 | 0.14 |
| **Macro Score** | 0.67 | 0.62 | 0.64 | 0.55 | 0.37 | 0.40 |
| **Weighted Score** | 0.70 | 0.70 | 0.70 | 0.65 | 0.71 | 0.64 |

## Acknowledgments

We are very thankful for the Google Colaboratory open-access server to perform these experiments.

## References

[1] C. Myers-Scotton, Common and uncommon ground: Social and structural factors in codeswitching, Language in society (1993) 475–503.

[2] 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.

[3] 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: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

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

[5] K. Rudra, S. Rijhwani, R. Begum, K. Bali, M. Choudhury, N. Ganguly, Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter?, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1131–1141. URL: https://www.aclweb.org/anthology/D16-1121. doi:10.18653/v1/D16-1121.

[6] A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Shrivastava, R. Mamidi, D. M. Sharma, Shallow parsing pipeline - Hindi-English code-mixed social media text, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1340–1345. URL: https://www.aclweb.org/anthology/N16-1159. doi:10.18653/v1/N16-1159.

[7] 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.

[8] 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.

[9] O. Habimana, Y. Li, R. Li, X. Gu, G. Yu, Sentiment analysis using deep learning approaches: an overview, Science China Information Sciences 63 (2020) 1–36.

[10] S. Sharma, P. Srinivas, R. C. Balabantaray, Text normalization of code mix and sentiment analysis, in: 2015 international conference on advances in computing, communications and informatics (ICACCI), IEEE, 2015, pp. 1468–1473.

[11] A. Pravalika, V. Oza, N. Meghana, S. S. Kamath, Domain-specific sentiment analysis approaches for code-mixed social network data, in: 2017 8th international conference on computing, communication and networking technologies (ICCCNT), IEEE, 2017, pp. 1–6.

[12] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

[13] K. Sarkar, Ju_ks@ sail_codemixed-2017: Sentiment analysis for Indian code mixed social media texts, arXiv preprint arXiv:1802.05737 (2018).

[14] G. Remmiya Devi, P. Veena, M. Anand Kumar, K. Soman, Amrita-cen@ fire 2016: Code-mix entity extraction for Hindi-English and Tamil-English tweets, in: CEUR workshop proceedings, volume 1737, 2016, pp. 304–308.

[15] E. Kouloumpis, T. Wilson, J. Moore, Twitter sentiment analysis: The good the bad and the omg!, in: Fifth International AAAI conference on weblogs and social media, Citeseer, 2011.

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

[17] A. Joshi, A. Prabhu, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2482–2491. URL: https://www.aclweb.org/anthology/C16-1234.

[18] S. Mukherjee, Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features, in: 2019 IEEE 16th India Council International Conference (INDICON), IEEE, 2019, pp. 1–4.

[19] I. TOMEK, Two modifications of CNN, IEEE Transactions on Systems, Man, and Cybernetics SMC-6 (1976) 769–772.

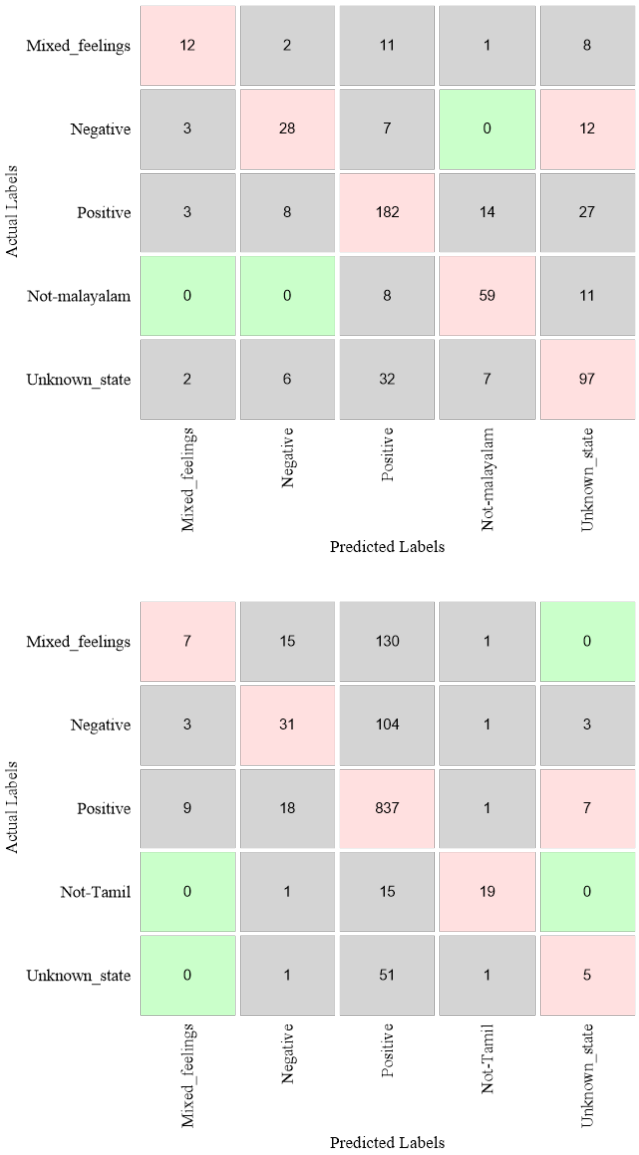


**Figure 2:** Confusion Matrix obtained on the validation set of Malayalam-English (upper) and Tamil-English (lower) code-mixed datasets

# A. Confusion Matrix