An ensemble model for sentiment classification on code-mixed data in Dravidian Languages

S R Mithun Kumar[1,2], Nihal Reddy[2], Aruna Malapati[2] and Lov Kumar[2]

[1]*Uber Research and Development India, Bangalore, India*
[2]*BITS Pilani, Hyderabad, India*

Abstract

Dravidian languages, Tamil, Kannada, Malayalam and Telugu, is spoken by over 220 million but is vastly under-resourced for natural language processing tasks. Code-switching and code-mixing have been on the rise, with multilingual speakers opting for expressing their opinion in their mother tongue along with English in both written text as well as in speech. Challenges arise in sentiment analysis of code-switched Dravidian languages because of under-resourced corpora and randomness in language interspersing. This paper applied an ensemble sentiment classification strategy based on majority voting using 13 different classification models on the Dravidian code-mixed languages dataset provided in FIRE 2021[1]. The key conclusion from our experiments was that the ensemble of multiple classifiers outperformed others for sentiment classification. Our approaches show that a result of weighted F1-score of 0.59, 0.65 and 0.60, respectively, on Kannada, Malayalam and Tamil code-switched data can be achieved with the traditional machine learning algorithms through an ensemble of multiple classifiers.

Keywords

Code-Mixing, Code-Switching, Dravidian, Tanglish, Kanglish, Manglish, Sentiment Classification

1. Introduction

Social media evolution has thrown numerous challenges and opportunities to the research community, and one such challenge is code-switching or code-mixing. Linguistic code-switching or code-mixing is the mixing of two or more languages in a conversation or even an utterance. While the majority of the current research has been on English mixed with languages like Spanish, or Indian languages like Tamil, Telugu, Hindi, Malayalam, Kannada in the Asian subcontinent, it could also extend to French and Arabic mixed in the African subcontinent or to trilingual code-switching among Cantonese, English and Putonghua (Chan 2019). Code-switched language is on the rise to as high as 20% in multilingual societies (Choudhury et al. 2019).

Speech assistants like Siri and Alexa have a deep interest in code-mixing since most users in multilingual society tend to use it regularly. It has multiple applications ranging from comments

[1]https://dravidian-codemix.github.io/2021/datasets.html

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

mithunkumar.sr@gmail.com (S. R. M. Kumar); h20201030161@hyderabad.bits-pilani.ac.in (N. Reddy); arunam@hyderabad.bits-pilani.ac.in (A. Malapati); lovkumar@hyderabad.bits-pilani.ac.in (L. Kumar)
0000-0002-1152-143X (S. R. M. Kumar); 0000-0002-7890-9850 (N. Reddy); 0000-0001-7275-378X (A. Malapati); 0000-0002-0123-7822 (L. Kumar)

on social media sites to day-to-day usage in written communication. Negative sentiments are more often expressed in mother-tongue than the positive sentiments, which are generally expressed in English making it necessary to learn the code-switched languages.

While monolingual NLP tasks form the basis and are no different to code-mixed languages in most of the aspects, significant challenges for code-mixed data exist in language identification, data collection and preparation strategy, optimally using the existing resources and on the user-centric design of code-mixed NLP systems. This amplifies even more when one of the languages is under-resourced.

Dravidian languages are vastly under-resourced, and when code-mixed with English is a harder NLP task. Sentiment analysis on code-switched Dravidian languages is still ongoing research which will help analyse the emotion and attitude of the users who express in code-switched languages with the rising usage on social media like TikTok, YouTube, Whatsapp, etc.

2. Related Work

Computational approaches to code-switching, related workshop and ACL anthology[1] has seen an increase in the research papers in the last three to four years.

Graph Convolutional Networks with multi-headed attention was experimented by Dowlagar et al. 2021 where it yielded a weighted F1-score of 0.75 for Malayalam-English code-mixed data with FIRE 2020 dataset published by Chakravarthi, Jose, et al. (2020).

Ensemble of character-trigrams based Long Short Term Memory (LSTM) model and word n-grams based Multinomial Naive Bayes (MNB) has been proposed by Jhanwar et al. (2018) for Hindi-English code-mixed language pair (Prabhu et al. 2016). This model takes in the combined strength of LSTM and probabilistic models. LSTM was performing better on longer length sentences due to its ability to capture sequential information whereas MNB performed generalisation on rare words.

All prior research highlighted above focus on deep learning techniques, which perform significantly well with longer length sentences. For instance, Jhanwar et al. (2018) experimented with datasets which has an average of fifty words. However, most social media content, such as Youtube comments, tend to be shorter. For instance Kannada code-mixed dataset of FIRE 2021 (Hande et al. 2020) has an average comment length of fewer than seven words. We argue that probabilistic and deterministic classifiers and an ensemble of traditional classifiers will yield the same or better results on datasets with shorter sentences.

Our approach was to build a pipeline with traditional classifiers, evaluate the performance metrics for sentiment classification and then iterate with ensemble techniques that could be set as a baseline for any short-length code-mixed text. This was done as part of the shared task on sentiment detection along the lines as described by Priyadharshini et al. (2021).

[1] https://aclanthology.org/search/?q=code+mixing

Table 1
Distribution of data in the Dravidian-CodeMix-FIRE 2021 dataset

Dataset	Positive	Negative	Mixed Feelings	Unknown State	Not the target language
Tamil	20,070	4,271	4,020	5,628	1,667
Malayalam	6,421	2,105	926	5,279	1,157
Kannada	2,823	1,188	574	711	916

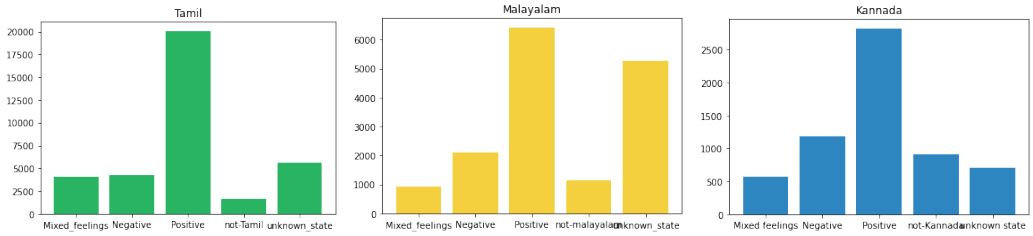

Figure 1: Bar plot on dataset split for Tamil, Malayalam and Kannada

3. Data

This section presents the detailed description of the dataset and its distribution, along with the research framework used.

3.1. Data Description

The dataset used for the task is from the official datasets released in Dravidian-CodeMix - FIRE 2021 which comprises labelled sentiment data of YouTube video comments on language pairs like Kannada-English (Hande et al. 2020), Malayalam-English (Chakravarthi, Jose, et al. 2020) and Tamil-English (Chakravarthi, Muralidaran, et al. 2020). The data has been code-switched language pairs, mostly in Roman script, both for English and the Dravidian language, wherein the latter has been transliterated from the source language to Roman script. However, there remains a good portion of the Dravidian script too in the data.

3.2. Data Distribution

The data distribution is shown in Table 1. The dataset contains labelled code-mixed sentences into five categories: Positive, Negative, Mixed Feelings, Unknown State and not in the intended language. The dataset contains inter-sentential, intra-sentential code-mixed sentences in Tamil, Malayalam and Kannada with English. As seen in Figure 1, the data is imbalanced, with most of the labels being available for the positive sentiment.

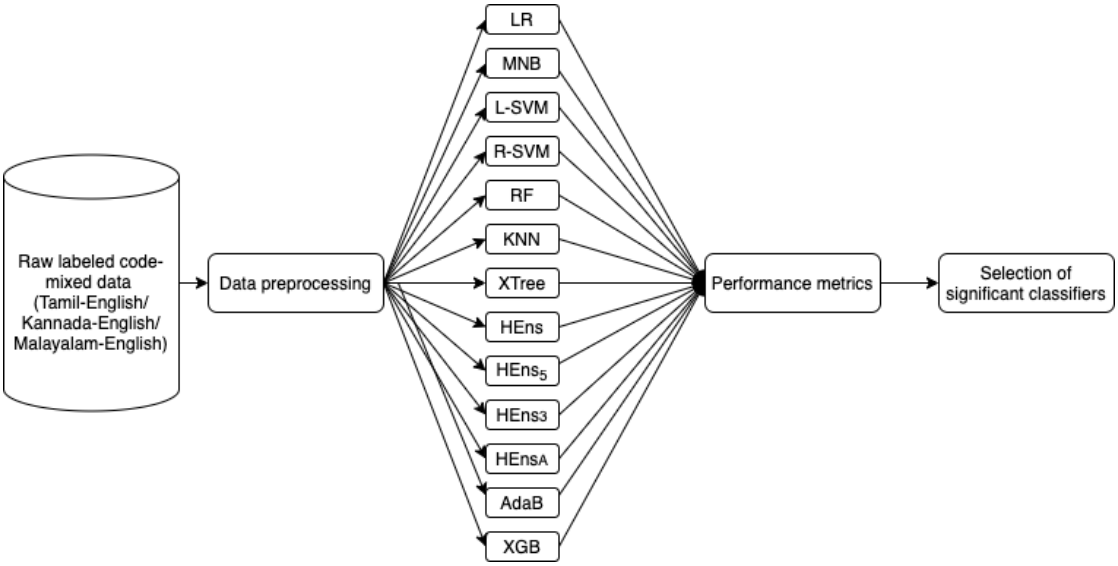

Figure 2: Framework representing the experimentation setting

4. Methodology

4.1. Data Preprocessing

The data has been preprocessed for removing stopwords, punctuation and emoticons. NLTK[2] library has been used for stemming, lemmatisation and removing stop-words. We have used the spaCy[3] library for named entity recognition.

4.2. Experiment Setting

The pipeline was set up to train the data, both on traditional as well as on ensemble techniques, as represented in Figure 2.

4.2.1. Traditional classifiers

In the first approach, the data has been run on multiple traditional machine learning algorithms for classification. The following parameters are common to all. 'vect', CountVectorizer, min_df=3, max_df=0.2, analyzer='word', ngram_range=(1, 3), TfidfTransformer().

This data was trained on traditional classifiers, including Logistic Regression (LR), Multinomial Naive Bayes (MNB), Linear SVM (L-SVM), RBF SVM (R-SVM), Poly SVM (P-SVM), Random Forest (RF), K-Nearest Neighbours (KNN) and Extra Tree Classifier (XTree).

[2]https://www.nltk.org/
[3]https://spacy.io/

4.2.2. Ensemble of multiple classifiers

The data was then run with ensemble classifiers with estimators as detailed out in Table 2. Various ensemble methods experimented with were AdaBoost (AdaB), XGBoost (XGB), Hard Ensemble of Voting Classifier (HEns), Hard Ensemble of Top 5, Top 3 and All Classifiers (HTop_5, HTop_3, HTop_A).

Table 2
Ensemble methods and the Estimators used

Hard Ensemble Method	Estimators
Voting Classifier (HEns)	L-SVM, LR, MNB, XTree, RF, P-SVM, R-SVM, KNN
Top 5 Classifiers (HTop_5)	L-SVM ,LR, RF, P-SVM, R-SVM
Top 3 Classifiers(HTop_3)	L-SVM, LR, P-SVM

5. Results

In this paper, eight different types of traditional machine learning and three different types of ensemble methods have been used to develop a sentiment prediction model for code-mixed language pairs in Tamil-English, Malayalam-English and Kannada-English. The predictive power of these sentiment prediction models are validated using 5-fold cross-validation and compared with four different performance metrics such as Precision, Recall, F1-Score and Accuracy. The performance values of these models are presented in Table 3 through which we derived the following observations:

- Ensemble classifiers generally outperformed all the single classifiers across all the three language pairs.
- The ensemble of a mix of both weak and strong individual learners always had a probabilistic model like logistic regression.
- Both, logistic regression model as well as an ensemble model, performed very close to each other.

5.1. Comparative Analysis: Box plot

In this work, box plot diagram for different performance metrics, precision, recall and F1-scores has been used to compare the performance of the developed models using different techniques. Figure 3 shows the box plot for each performance metric, precision, recall and F1-scores compared with each classifier. The information present in Figure 3 suggested that the ensemble methods generally perform better than other classifier. The information present in Figure 3 also suggested that the probabilistic models like logistic regression, in silo, perform better than any other stand-alone classifiers. This performs even better with an ensemble of top classifiers. The performance metrics are very close to the values observed in the baseline model using transformer-based models on the FIRE 2021 dataset published by Chakravarthi,

Priyadharshini, Muralidaran, et al. (2021), which achieved an F1-score of 0.67, 0.59 and 0.62 for Kannada, Malayalam and Tamil code-mixed datasets, respectively as published by Chakravarthi, Priyadharshini, Thavareesan, et al. (2021). Our experimentation with ensemble models shows that the best scores of 0.59, 0.67 and 0.60 can be achieved for the same set of language pairs.

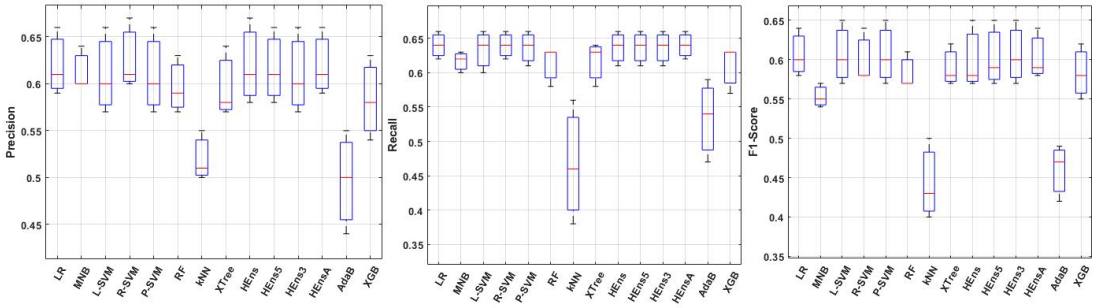

Figure 3: Box plot of Precision, Recall and F1 scores of all the methods

5.2. Comparative Analysis: T-test

In this work, the T-test technique has also been applied to find the significant difference in the performance of the developed models using different classifiers. The T-test is used to test our considered null hypothesis, i.e., "There is no significant difference in the performance of the developed sentiment prediction models using different techniques". Figure 4 shows the

Table 3
A comparison table of results from classifying using different methods. Bold values are the best F1 scores.

Method	Kannada				Malayalam				Tamil			
	P_k	R_k	F1_k	A_k	P_m	R_m	F1_m	A_m	P_t	R_t	F1_t	A_t
LR	0.59	0.62	0.58	0.62	0.66	0.66	0.64	0.66	0.61	0.64	**0.60**	0.64
MNB	0.60	0.62	0.57	0.62	0.63	0.60	0.55	0.60	0.60	0.63	0.54	0.63
L-SVM	0.57	0.60	0.57	0.60	0.65	0.66	0.64	0.66	0.60	0.64	**0.60**	0.64
R-SVM	0.60	0.62	0.58	0.62	0.67	0.66	0.63	0.66	0.61	0.64	0.58	0.64
P-SVM	0.57	0.61	0.57	0.61	0.65	0.66	0.64	0.66	0.59	0.64	0.60	0.64
RF	0.54	0.56	0.55	0.56	0.63	0.64	0.61	0.64	0.58	0.63	0.57	0.63
KNN	0.55	0.38	0.40	0.38	0.51	0.46	0.43	0.46	0.50	0.56	0.50	0.56
XTree	0.56	0.58	0.57	0.58	0.64	0.65	0.63	0.65	0.59	0.63	0.58	0.63
HEns	0.59	0.62	0.58	0.62	0.67	0.67	**0.65**	0.67	0.61	0.64	0.58	0.64
HEns_5	0.58	0.61	0.57	0.61	0.66	0.66	**0.65**	0.66	0.61	0.64	0.59	0.64
HEns_3	0.57	0.61	0.57	0.61	0.65	0.66	0.64	0.66	0.60	0.64	**0.60**	0.64
HEns_A	0.60	0.62	**0.59**	0.62	0.66	0.66	0.64	0.66	0.61	0.64	0.59	0.64
AdaB	0.55	0.54	0.47	0.54	0.50	0.47	0.42	0.47	0.44	0.59	0.49	0.59
XGB	0.53	0.57	0.55	0.57	0.63	0.63	0.62	0.63	0.58	0.63	0.58	0.63

result of different techniques for each of the performance metric, precision, recall and F1 scores. The green dots in Figure 4 indicate that the considered null hypothesis is accepted, i.e., the performance of the models does not depend on techniques; similarly, red dots indicate that there would be a difference in the performance of the models developed using different techniques. From Figure 4, we can observe that the predictive power of the models developed using different techniques are significantly different. From Figure 4, we can also observe that the ensemble methods significantly improved the performance of the models.

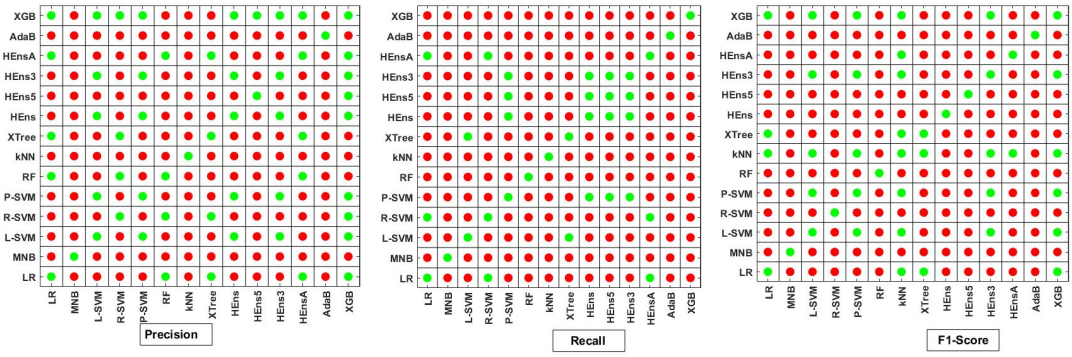

Figure 4: T-test: Precision, Recall and F1 scores of all the methods

6. Conclusion

In this work, we applied different traditional machine learning methods as well as ensemble methods on code-mixed data from FIRE 2021, which had Tamil-English, Malayalam-English and Kannada-English language pairs with an objective to develop sentiment prediction models. The performance of these developed sentiment prediction models are computed using precision, recall and F1-score. Our experimental results show that:

- The proposed ensemble classifier performs better than any stand-alone classifier.
- The developed models based on ensemble technique achieved an F1-score of 0.59 and accuracy of 0.62 for Kannada.
- The developed models based on ensemble technique achieved an F1-Score of 0.65 and an accuracy of 0.67 for Malayalam.
- The developed models based on ensemble technique achieved an F1-score of 0.60 and an accuracy of 0.64 for Tamil.

The future steps would be to better the results through the transliteration-translation task to augment the preprocessing and complete the sentiment analysis on the monolingual English corpus, rather than a bilingual corpus for code-switched languages.

7. Acknowledgement

Thanks to Dr Aravind Ranganathan, Uber R&D, and the anonymous reviewers for the valuable suggestions and thorough review comments.

8. Appendix

Our code is available on GitHub[4]

References

Chakravarthi, Bharathi Raja, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae (May 2020). "A Sentiment Analysis Dataset for Code-Mixed Malayalam-English". 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)*. Marseille, France: European Language Resources association, pp. 177–184. ISBN: 979-10-95546-35-1. url: https://www.aclweb.org/anthology/2020.sltu-1.25.

Chakravarthi, Bharathi Raja, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae (May 2020). "Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text". 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)*. Marseille, France: European Language Resources association, pp. 202–210. ISBN: 979-10-95546-35-1. url: https://www.aclweb.org/anthology/2020.sltu-1.28.

Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae (2021). *DravidianCodeMix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-Mixed Text*. arXiv: 2106.09460 [cs.CL].

Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Sajeetha Thavareesan, et al. (2021). "Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text". In: *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. Online: CEUR.

Chan, Ka Long Roy (2019). "Trilingual Code-switching in Hong Kong". In: *ALRJournal* 3.4, pp. 1–14. DOI: 10.14744/alrj.2019.22932.

Choudhury, Monojit, Anirudh Srinivasan, and Sandipan Dandapat (Nov. 2019). "Processing and Understanding Mixed Language Data". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*. Hong Kong, China: Association for Computational Linguistics. url: https://aclanthology.org/D19-2002.

Dowlagar, Suman and Radhika Mamidi (Apr. 2021). "Graph Convolutional Networks with Multi-headed Attention for Code-Mixed Sentiment Analysis". In: *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Kyiv: Association for Computational Linguistics, pp. 65–72. url: https://aclanthology.org/2021.dravidianlangtech-1.8.

[4]https://github.com/mithunkumarsr/CodeMixingDravidianLanguage

Hande, Adeep, Ruba Priyadharshini, and Bharathi Raja Chakravarthi (Dec. 2020). "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*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 54–63. url: https://www.aclweb.org/anthology/2020.peoples-1.6.

Jhanwar, Madan Gopal and Arpita Das (2018). "An Ensemble Model for Sentiment Analysis of Hindi-English Code-Mixed Data". In: *CoRR* abs/1806.04450. arXiv: 1806.04450. url: http://arxiv.org/abs/1806.04450.

Prabhu, Ameya, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma (2016). *Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text*. arXiv: 1611.00472 [cs.CL].

Priyadharshini, Ruba, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy (2021). "Overview of the DravidianCodeMix 2021 Shared Task on Sentiment Detection in Tamil, Malayalam, and Kannada". In: *Forum for Information Retrieval Evaluation*. FIRE 2021. Online: Association for Computing Machinery.