

Voting ensemble model based Malayalam-English sentiment analysis on code-mixed data

K.Nimmi , B.Janet

Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India

Abstract

The most serious issue with code-mixing is that people switch between languages (for example, Malayalam and English) and type in English instead of writing Malayalam words. Traditional NLP models can't handle data code-mixing data. Sentiment Analysis on Kannada-English, Malayalam-English, or Tamil-English code-mixed datasets based on five labels is the Dravidian Code-Mixed FIRE 2021 challenge. The classification is to be done based on the following labels 'Not-Malayalam, "Neutral state, 'Positive, 'Mixed feelings, 'Negative'. This paper focuses on Malayalam-English code-mixed data sentiment analysis based on the Ensemble voting model with machine learning models - Support Vector machine (SVM) and Logistic Regression and Bagging. The Hard Voting classifier model provided an accuracy : 67.78 % and F1-score : 67.53%.

Keywords


Voting ensemble model, Bagging, Code mixed, Support Vector Machine, Logistic regression.

1. Introduction

Code-Mixing (CM) is a common occurrence in which different languages are used in a single sentence or multiple sentences. CM is primarily used in casual discussions and on different social media sites. Sentiment Analysis (SA) is a key stage in Natural Language Processing (NLP) that has been extensively researched in monolingual texts. Because of it's non-standard representations, code-mixing makes sentiment analysis more difficult. When creating text, code-mixing refers to the blending of languages. The most serious issue with Malayalam-English code-mixing is that individuals switch between languages (for example, English and Malayalam) and type in English phonetically while writing Malayalam words in Dravidian scripts. Traditional NLP models are trained on large monolingual datasets (for example, Malayalam or English); code-mixing is difficult since traditional NLP models cannot manage code-mixing. In most cases, code-mixing entails combining two languages to produce a new language that combines parts from both in a structurally comprehensible way. Bilingual and multilingual communities have been seen to utilize several languages in casual speech and communication.


Code mixing (CM) [1] refers to combining an utterance of another language with linguistic units of languages, according to Myers-Scotton. Not everyone is comfortable with a single language, some people from various linguistic origins and cultures express their feelings on a subject in mixed language [2]. The accidental switching between different languages in the

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

 nimmi.nitt@gmail.com (K.Nimmi); janet@nitt.edu (B.Janet)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

same discussion is code mixing [3]. Words from various languages can be found in codemixed languages. According to Choudhary et al. Code-mixed [4] data borrows vocabulary and syntax from various languages, and it frequently creates new structures based on user input.

During sentiment analysis the sentiments are extracted from the dataset and categorize them for usage in a variety of applications such as customer service, social media content moderating and reputation management, [5]. Based on feedback or even the polarity of remarks, sentiment analysis has benefited the industry in generating a summary of human opinions and interests, [6]. This paper focuses on (Malayalam-English) code mixed comments classification using an Ensemble voting model formed using Support Vector Machine (SVM), Logistic Regression and Bagging. The performance of hard voting was higher when compared to the soft voting model.

2. Related Works

Advani et al. proposed a machine learning algorithm that can detect the difference between positive and negative feelings based on lexical, metadata features and sentiment[7]. Sharma et al. developed a shallow Hindi-English code-mixed social media text parser [8]; this shallow parser model was modelled as three separate sequence labeling problems. Singh and Lefever developed a cross lingual embeddings technique for Hinglish code mix data [9] that is unsupervised. The performance of various transformer models is analyze in [10] on code-mixed data sentiment analysis. To identify sentiment on code mixed text (Hindi and English), [11] used a lexicon-based technique. In a code-mixed Hinglish dataset, [12] shows a strategy for detecting hate speech. Only a few research have used code-mixed Dravidian language datasets for sentiment analysis the details of text classification using deep learning models to identify the Malayalam-English [13] and Tamil-English sentiments are provided in [14] ; Dravidian-CodeMix-FIRE 2020 was a sentiment polarity classification challenge on code mix data for classification of Youtube into five classes which is based on the code mix dataset. A computational technique was proposed by Das and Bandyopadhyay used publicly available English Sentiment lexicons and a bilingual English-Bengali dictionary [15] to create a SentiWordNet equivalent for the language Bengali. Baruah et al. found that on code-mixed Malayalam text, the SVM classifier [16] trained using TF-IDF word and character features of n-grams performed the best.

3. Methodology

3.1. Data and Pre-Processing

The Dravidian Code-Mix FIRE 2021 project's purpose is to categorize YouTube comment code-mixed datasets written in Kannada-English, Malayalam-English, or Tamil-English into five labels based on sentiment polarity. Emotion polarity is assigned to each comment at the comment level. In real-world situations, this dataset has concerns with class imbalance. In Dravidian-CodeMix-FIRE 2021 challenge more dataset was provided with the majority of the comments written in either Malayalam, Tamil, or Kannada grammar with English lexicon or Malayalam, Tamil, or Kannada lexicon with English grammar. Variety of comments were written in Malayalam, Tamil, and Kannada script, with English phrases interleaved. The detailed description of datasets used

is provided in two papers [17] and [18] along with details on how the data was acquired and labeled in the datasets. In Table 1, the statistics for each class are tabulated.

Table 1

Statistics of training , validation and testing set.

Label	Train	Valid	Test
Positive	6421	706	780
unknown_state	5279	580	643
Negative	2105	237	258
not-malayalam	1157	141	147
Mixed_feelings	926	102	134
Total	15,889	1,767	1,963

- Positive: The language contains an explicit or implicit hint that the speaker is in a positive mood, such as admiring, joyful, forgiving, and calm.
- Negative: The language contains a clear or tacit hint that the speaker is in a negative mood, such as angry, sad, aggressive, or worried.
- Mixed feelings: The language contains an explicit or implicit hint that the speaker is experiencing both the above emotions.
- Neutral: No indication of the speaker’s emotional condition is provided, either explicitly or implicitly.
- Non-Malayalam: The language isn’t what it’s supposed to be: If a sentence does not contain any words Malayalam, it is not Malayalam.

We employed pre-processing methods to clean the comments in the dataset, such as replacing emojis with similar words, changing the comment to lower case, eliminating stop words, HTML tags, and accented characters.

3.2. Model Description

Ensemble learning can take several forms, such as stacking, bagging, and so on. Because of the bagging method’s efficiency and ease, this paper created an ensemble model as both hard voting and soft voting classifier based on bagging by combining three classifiers. Conventional Machine Learning classifier; Support Vector Machine (SVM) [19], bagging model and Logistic Regression [20] are combined to create an ensemble model for Dravidian code mix sentiment identification. The goal of ensembling of simple classifiers is to create a robust classifier that takes advantage of each classifier’s strengths. TF-IDF [21] vectors obtained during feature engineering are used to train the model. Figure1 describes the model block diagram.

4. Implementation Details

We used a Windows 10, 64-bit operating system, Intel Core i5 CPU 2.40 GHz, and IDE drive for code-mix data sentiment analysis. We employ Anaconda 2019.10, an open-source program

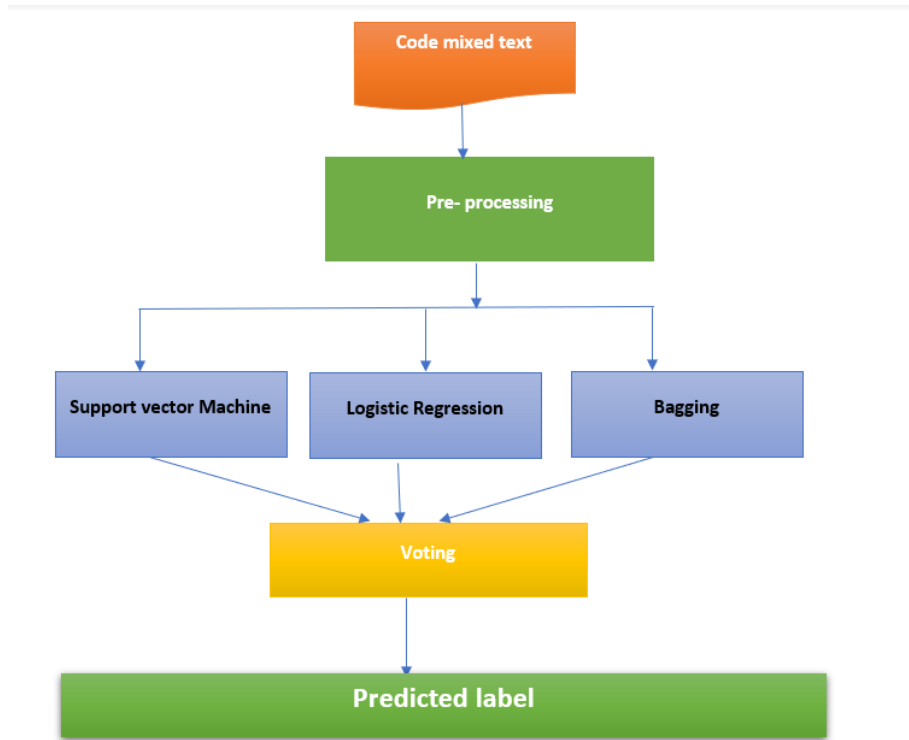


Figure 1: Block diagram of Voting model used for code-mix data sentiment analysis

for developing Machine Learning projects and current Python libraries such as scikit-learn and Pandas (IDE). To extract features, the tf-idf technique [21] is used. Various parameters were tried, but the best results were provided by the following parameters. For SVM model, radial basis function kernel ('rbf') is used with a 'C' value of 100. The 'c' value used in logistic regression is 100, and the kernel is 'linear.' The bagging model used in this experiment use SVM as a base estimator. The ensemble model used for classification is majority voting. We had trained the model for Malayalam English code-mix data. Sklearn library [22] is used to build the model.

5. Results

Hard Voting classifier provided an accuracy of 67.78 % and precision of 67.17% and recall of 67.78% and F1-score of 67.04%. Soft voting classifier have provided an Accuracy of 67.53% and precision of 66.93% and Recall of 67.53% and F1-score of 66.96% respectively. In Table 2 the classification metrics of Hard Voting classifier is provided and in Table 3 the classification metrics of soft ensemble model on test data is provided. The overall model performance is provide in the Table 4. From the Table 4, it is clear that Hard Voting model achieved slightly

better results when compared to the Soft Voting model. The label positive and unknown_state was predicted with higher precision, recall and F1-score than other labels because of the large number of the training set. The voting model performance is hampered by class imbalance.

Table 2

Performance evaluation metrics of Voting ensemble (Hard).

Label	Precision	Recall	f1-score
Positive	0.69	0.79	0.74
unknown_state	0.71	0.68	0.69
Negative	0.60	0.48	0.53
Not-Malayalam	0.68	0.73	0.71
Mixed_feelings	0.54	0.32	0.40

Table 3

Performance evaluation metrics of Voting ensemble (Soft)

Label	Precision	Recall	f1-score
Positive	0.70	0.79	0.74
unknown_state	0.71	0.67	0.69
Negative	0.58	0.50	0.54
Not-Malayalam	0.67	0.71	0.69
Mixed_feelings	0.49	0.34	0.41

6. Conclusion

This paper uses a Hard and Soft voting ensemble model to categorize Code-Mixed Malayalam and English comments dataset. Hard Voting ensemble model provided slightly better accuracy than Soft Voting classifier. In the future, applying a deep learning ensemble model on the dataset can improved the model performance and fine-tuning the model to find sarcastic Manglish comments.

Table 4

Total performance of the model

Performance measures	Voting Classifier (Hard)	Voting Classifier (Soft)
Accuracy	67.78%	67.53%
Precision	67.17%	66.93%
Recall	67.78%	67.53%
F1-Score	67.04%	66.96%

7. Opportunities in code-mixing

The demand for sentiment analysis of text from social media, which is usually code-mixed, is growing. However, there are limited tools available to develop models specifically for code-mixed data. The papers which provide details of codemix data are listed below. The research in [23] provide details of a dataset of Dravidian languages (Kannada, Malayalam, and Tamil) for detecting abusive language in social media and filtering user-generated material in local languages. In the study [24], the authors introduce a Kannada CodeMixed Dataset (KanCMD), a code mixed text, multi-task learning dataset in Kannada language for offensive language detection and sentiment analysis. The paper [25] describe a resource named (TamilMemes) which is used to recognizing a troll meme in Tamil, which help in detecting and dealing with trolls so that before harming an individual. Chakravarthi et al. explain how they created the corpus and assigned polarities to code-mixed Malayalam-English code-mixed data[26] and Chakravarthi et al. provides details of Tamil-English [27], respectively for sentiment analysis using comments on social media, thereby providing a benchmark dataset.

References

- [1] C. Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*, Oxford University Press, 1997.
- [2] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, P. Buitelaar, A dataset for troll classification of tamilmemes, in: *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, 2020, pp. 7–13.
- [3] B. R. Chakravarthi, M. Arcan, J. P. McCrae, 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, 2019.
- [4] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of code-mixed languages leveraging resource rich languages, *arXiv preprint arXiv:1804.00806* (2018).
- [5] S. Thavareesan, S. Mahesan, Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation, in: *2019 14th Conference on Industrial and Information Systems (ICIIS)*, IEEE, 2019, pp. 320–325.
- [6] S. Thavareesan, S. Mahesan, Word embedding-based part of speech tagging in tamil texts, in: *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, IEEE, 2020, pp. 478–482.
- [7] 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).
- [8] A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, D. M. Sharma, Shallow parsing pipeline for hindi-english code-mixed social media text, *arXiv preprint arXiv:1604.03136* (2016).
- [9] P. Singh, E. Lefever, Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings, in: *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, 2020, pp. 45–51.

- [10] S. Banerjee, B. R. Chakravarthi, J. P. McCrae, Comparison of pretrained embeddings to identify hate speech in indian code-mixed text, in: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, 2020, pp. 21–25.
- [11] 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.
- [12] P. Rani, S. Suryawanshi, K. Goswami, B. R. Chakravarthi, T. Fransen, J. P. McCrae, A comparative study of different state-of-the-art hate speech detection methods in hindi-english code-mixed data, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 42–48.
- [13] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed malayalam-english, arXiv preprint arXiv:2006.00210 (2020).
- [14] 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: Forum for Information Retrieval Evaluation, 2020, pp. 21–24.
- [15] A. Das, S. Bandyopadhyay, Subjectivity detection in english and bengali: A crf-based approach, Proceeding of ICON (2009).
- [16] A. Baruah, K. A. Das, F. A. Barbhuiya, K. Dey, Iitg-adbu@ hasoc-dravidian-codemix-fire2020: Offensive content detection in code-mixed dravidian text, arXiv preprint arXiv:2107.14336 (2021).
- [17] 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.
- [18] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, T. Durairaj, 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.
- [19] M. Ahmad, S. Aftab, I. Ali, Sentiment analysis of tweets using svm, Int. J. Comput. Appl 177 (2017) 25–29.
- [20] R. E. Wright, Logistic regression. (1995).
- [21] T. Tokunaga, I. Makoto, Text categorization based on weighted inverse document frequency, in: Special Interest Groups and Information Process Society of Japan (SIG-IPJS), Citeseer, 1994.
- [22] T. P. Trappenberg, Machine learning with sklearn, in: Fundamentals of Machine Learning, Oxford University Press, 2019, pp. 38–65.
- [23] B. R. Chakravarthi, R. Priyadharshini, N. Jose, A. Kumar M, T. Mandl, P. K. Kumaresan, R. Ponnusamy, H. R L, J. P. McCrae, E. Sherly, Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 133–145. URL: <https://aclanthology.org/2021.dravidianlangtech-1.17>.
- [24] 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://aclanthology.org/2020.peoples-1.6>.

- [25] S. Suryawanshi, B. R. Chakravarthi, 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, Association for Computational Linguistics, Kyiv, 2021, pp. 126–132. URL: <https://aclanthology.org/2021.dravidianlangtech-1.16>.
- [26] 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://aclanthology.org/2020.sltu-1.25>.
- [27] 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://aclanthology.org/2020.sltu-1.28>.