Machine learning based approach for sentiment Analysis on Multilingual Code Mixing Text

B Bharathi[1], G U Samyuktha[2]

[1]*Department of CSE, Sri Siva Subramaniya Nadar College of Engineering,Tamil Nadu, India*
[2] *Department of CSE, AVC College of Engineering, Mayiladuthurai*

Abstract

Social networking as become the irreplaceable platform like never before. We are more up to date with the issues whether they are good or bad around the globe. The oversharing happening on social media leads to cyberbullying. In this study we are going to compare and analyze methods for comment-level text polarity classification task using the Dravidian-CodeMix-FIRE2021 data-set. Techniques such as TFIDF, Count vectorizer and multilingual transformer based encoded features. The features are trained with different machine learning models such as Multi layer perceptron, SVM, Random forest. Our models scored F1 scores of 0.588, 0.69 and 0.63 for the Tamil-English, Kannada-English and the Malayalam-English code-mixed test data respectively.

Keywords

Sentiment analysis, Dravidian languages, Transformer, Machine learning approach

1. Introduction

With the advancement of technology, the era of meaningful information from social media data has arrived. Traditionally, sentiment analysis is done in text, but now a large amount of data is loaded such as views, images, emoticons and videos. By checking these data, we can analyze, verify and discover the public's sentiment towards specific events. Over the years, people have believed that the emoji is a means of communication, used in text or simply to express their emotions effectively. As the native language usage in social media increases, it is important to construct models which handles the combination of native language mixed with English language in the text[1],[2]. This paper proposes the machine learning approaches for Dravidian languages using the dataset provided in Dravidian-Code Mix-FIRE2021[3]. Dravidian code-mixed languages, including Malayalam, Kannada, and Tamil, are increasingly used by many people in social media [4][5]. The language is commonly written in Roman script. With the rise in the number of non-English and multilingual speakers using social media, there is an interest in analyzing the sentiment of the content posted by them. As code-mixed data does not belong to one language and is often written using a Roman script, identifying its polarity cannot be done using traditional sentiment analysis models.

FIRE 2021: Forum for Information Retrival Evaluation, December 13-17, 2021, India
bharathib@ssn.edu.in (B. Bharathi); gusamyuktha@gmail.com (G. U. Samyuktha)
https://www.ssn.edu.in/staff-members/dr-b-bharathi/ (B. Bharathi)
0000-0001-7279-5357 (B. Bharathi)

Table 1
Dataset Distribution

Data set description	No. of training sentences	No. of validation sentences	No. of test sentences
Tamil-English	35,657	3,963	4,403
Malayalam-English	15,889	1,767	1,963
Kannada-English	6213	692	768

The paper is organized as follows: The section 2, explains work related with sentiment analysis. The dataset descriptions are given in Section3.1 Section 3.2 details the experimental setup and various features used for this task. Section 4 provides a subjective analysis and comparison of the performance of various models on the development and test data. Finally, Section 5concludes the paper.

2. Related work

In the recent years,Sentiment analysis on multilingual code mixing text is an active research area [6]. For Kannada - English code mixed sentiment analysis text, distributed representation is used[7]. They have compared their results with different machine learning and deep learning techniques. The authors [8] reported that machine learning and neural networks algorithms achieves better accuracy for code-mixed social media for the systems relying on hand-crafted features. For sentiment and offensive language detection for Kannada language KanCMD [9] dataset is used.

3. Proposed work

3.1. Dataset Description

This data set consists of YouTube comments with emotional polarity tags at the message level. The categories of assignments include positive, negative,mixed feelings, unknown-state or if comments are not in the language of the label. This becomes a multi-class classification task. The average sentence length contained in single comments is 1. But there is a certain imbalance in the class because it simulates real life scenarios. More details about the dataset is found in [5]and [4]. The data distribution is tabulated in Table 1.

3.2. Experiments

The experimental structure of the task can be divided into two stages: feature extraction stage and classifier stage. In the feature extraction stage, technologies such as count vectorization, TFIDF vectorization [10], multilingual transformer based encoding etc. were analyzed, and different classifiers such as support vector machine, logistic regression, multilayer perceptron, naive Bayes, Knearest and random forest classifier were compared. The features extracted

from the first stage are used to train the machine learning model in the second stage, and their performance is compared using the F1 score and the accuracy score. The measurement methods in the scikit learn package are used to measure performance. The code for the sentiment analysis task is available here [1]

3.2.1. Count and TFIDF Vectorization

The content of the text annotation is a mixture of various languages, their grammar and switching between different symbols. It becomes difficult to capture the consistent intensity of comments using existing pre-trained models. Therefore, the model of words and characters based on bag-of-words is realized and analyzed by changing the range of ngram. The ngram ranges of 2-3, 1-5 and 2-3 gave the better results for Tamil-English,Malayalam-English and Kannada-English corpus respectively. Subtask gives the best results in devset. The term frequency inverse document frequency model helps to assign a weight of less than to the mediocre words in the corpus. This technique emphasizes unique terms in the corpus more than repeated words and provides a better model.

3.2.2. Multilingual Embedding Models

The YouTube comments selected for the research contained text from the fused English and Dravidian languages. This becomes the main issue to consider when applying the mono language pre-training model and adjusting it for this specific task. However, by unsupervised selection of pre-trained models in a large number of languages, can fine-tune these multilingual models to fit well with Codemix applications. Since the multilingual model fastText and BERT [11] have shown fruitful results, is considered in this experiment. The sentence transformers such as paraphrase-xlm-r-multilingual-v1 ,stsb-xlm-r-multilingual, paraphrase-multilingual-mpnet-base-v2 encodes the sentence into 768 dimensional dense vector space [12]. These sentence transformers were used in Tamil-English, Malayalam-English code mixed text. The dimension for distiluse-base-multilingual-cased-v2 and distiluse-base-multilingual-cased-v2 is 512 used in Kannada-English code mixed text.

4. Performance analysis

On account of all the analyzed models the TFIDF gives the best performance for the Tamil-English code mixed text for development data, if we consider the Malayalam-English Corpus that TFIDF and countvectorizer techniques generated equal performance and if we consider Kannada-English Corpus TFIDF gives the best performance. The performance details for each model on development set is listed in the Tables 2,3, 4.

From Table 2, it has been noted that count vectorizer with MLP and SVM is giving equal performance using F1-score.The multilingual transformer model's performance is lower than TFIDF and count vectorizer features.

[1] https://github.com/bhassn/Fire21.git

Table 2
Performance of Tamil-English code-mixed data using dev-data

Features	Classifier	Precision	Recall	F1-score
Countvec	MLP	0.56	0.57	0.57
Countvec	Randomforest	0.62	0.62	0.53
Countvec	SVM	0.57	0.57	0.57
Countvec	Knearest	0.55	0.58	0.56
Tfidf	MLP	0.56	0.57	0.56
Tfidf	Randomforest	0.62	0.61	0.51
Tfidf	Naive Bayes	0.58	0.46	0.48
Tfidf	SVM	0.59	0.64	0.58
paraphrase-xlm-r-multilingual-v1	MLP	0.55	0.55	0.55
stsb-xlm-r-multilingual	MLP	0.52	0.54	0.53
paraphrase-multilingual-mpnet-base-v2	MLP	0.54	0.54	0.54

Table 3
Performance of Malayalam-English code-mixed data using dev-data

Features	Classifier	Precision	Recall	F1-score
Countvec	MLP	0.72	0.72	0.72
Countvec	Randomforest	0.72	0.68	0.65
Countvec	SVM	0.68	0.69	0.68
Tfidf	MLP	0.72	0.72	0.72
Tfidf	Randomforest	0.71	0.67	0.63
Tfidf	SVM	0.72	0.73	0.72
paraphrase-xlm-r-multilingual-v1	MLP	0.61	0.62	0.61
stsb-xlm-r-multilingual	MLP	0.59	0.59	0.59
paraphrase-multilingual-mpnet-base-v2	MLP	0.61	0.61	0.61

From Table 3, it has been noted that count vectorizer with MLP and TFIDF with MLP and SVM is giving equal performance using F1-score.The multilingual transformer model's performance is lower than TFIDF and count vectorizer features.

From Table 4, it has been noted that TFIDF with Random forest and SVM is giving better performance then other approaches for Kannada - English code mixed text.The multilingual transformer model's performance is lower than TFIDF and count vectorizer features.

The performance of code-mixed corpus using test data is tabulated in Table 5. The results submitted for this task where the best models stood 1st, 8th and 11th ranks in the Kannada, Malayalam and Tamil tasks respectively.

Table 4
Performance of Kannada-English code-mixed data using dev-data

Features	Classifier	Precision	Recall	F1-score
Countvec	MLP	0.60	0.60	0.60
Countvec	Randomforest	0.64	0.65	0.62
Countvec	SVM	0.60	0.59	0.60
Tfidf	MLP	0.61	0.62	0.61
Tfidf	Randomforest	0.66	0.66	0.63
Tfidf	SVM	0.68	0.68	0.65
paraphrase-xlm-r-multilingual-v1	MLP	0.58	0.59	0.58
distiluse-base-multilingual-cased-v2	MLP	0.59	0.61	0.60
distiluse-base-multilingual-cased-v1	MLP	0.60	0.61	0.60

Table 5
Performance of code-mixed data using test data

Data set	Precision	Recall	F1-score	Rank
Tamil-English	0.597	0.643	0.588	11
Malayalam-English	0.691	0.692	0.69	8
Kannada-English	0.639	0.656	0.63	1

5. Conclusion

This proposed work summarizes the machine learning techniques used for sentiment analysis in the last periods. The impact of applying data transformation can improve the implementation of the classification method, but the type of transformation depends on the dataset and the language it contains. Therefore, check the details, select characteristics, apply transformation and filter the least relevant data, generalize and the machine learning methods are effective, because the computers of today have limitations and cannot process all the data without previously processing any formal review. In this study, we have analyzed a variety of feature extraction techniques and conclude that the Count, TFIDF based vectorization, and multilingual transformer encoding technique performs well on code-mix polarity labeling task. With these features, we reach a weighted F1 score of 0.588 for the Tamil-English task, 0.69 for the Malayalam-English task and 0.63 for the Kannada-English tasks respectively.

References

[1] 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.

[2] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, P. B, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] K. Shalini, H. B. Ganesh, M. A. Kumar, K. P. Soman, Sentiment analysis for code-mixed indian social media text with distributed representation, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1126–1131. doi:10.1109/ICACCI.2018.8554835.

[8] P. Mishra, P. Danda, P. Dhakras, Code-mixed sentiment analysis using machine learning and neural network approaches, CoRR abs/1808.03299 (2018). URL: http://arxiv.org/abs/1808.03299. arXiv:1808.03299.

[9] 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.

[10] A. B. Nitin Nikamanth, B. Bharathi, J. Bhuvana, Ssncse_nlp@dravidian-codemix-fire2020:sentiment analysis for dravidian languages in code-mixed text, in: Working Notes of FIRE 2020- Forum for Information Retrieval Evaluation, CEUR, 2020, pp. 4–12.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.