Sentiment Analysis Model For Code-Mixed Tamil Language

N Sripriya [1] and S Divya [2]

[1,2] *Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India*

Abstract

Social Media is a vital source for communicating information and retrieval. To legitimize the contents in social media, sentiment analysis is vital and has become a most focused research area. Sentiment analysis is a Natural Language Processing (NLP) task and has been well analyzed for application in monolingual text. Sentiment analysis tasks become complex when applied to Code-mix data. Since the comments produced by viewers in social media incorporate emoticons and maybe in mixed language, sentimental analysis of such data is challenging. This paper describes a model that codes the input data by looking at the frequency of terms and is then categorized using a multiclass classification algorithm. This model is straightforward and produces better results in classifying the data based on the terms available in the input sequence. Evaluation of this model yields an average weighted F1 score of 0.35 is achieved when applied to the Dravidian Code-mix dataset produced for the Sentiment Analysis task in FIRE-2021.

Keywords

Sentiment, emoticons, Code-mix, Natural Language Processing.

1. Introduction

Sentiment analysis is a taxonomy task that is used to extract sentiments from text data. This task has its benefits in numerous applications like customer feedback, reputation management and legalizing content in social media [1], [2], [3]. This is widely used in generating a summary of human ideas or interests extracted from the comments posted by the users or viewers [4]. Many online forums allow users to share their experiences as product or content reviews. To facilitate the user, the online platforms ensure the mother tongue communication or Code-mix language to share the user's view in a realistic way. Since most of the MLP tasks are trained over well-organized data with proper grammar, it becomes challenging when being applied to user-generated comments [5].

Code-mixing or Code-switching alternates two or greater numbers of languages at various levels of the content. It may be done at a document level, paragraph level, comments level, sentence level, phrase level, word level, or at even morpheme level. This represents a unique way of conversing in a bilingual or multilingual society [6].

This paper elaborates a model that generates embedding representation for the text data available in the dataset issued for the sentiment analysis task by Dravidian Code-mix FIRE 2021. This is a multiclass classification problem that generates five different labels for the data collected from YouTube comments. The developed model extracts functionality from the given input data and based on those features the input data is classified into several classes. This classification is done using a Machine Learning algorithm, which learns from the features extracted and the labels given to each training data during the training stage. Based on the learning, it tries to classify the data into distinct groups and labels each data by the group it belongs to. Since the classification task tries to classify the data into multiple classes, multiclass classifiers are used for categorizing the given data.

FIRE 2021, Forum for Information Retrival Evaluation, December 13-17, 2021, India
E.MAIL: sripriyan@ssn.edu.in(N.Sripriya); divyas@ssn.edu.in (S.Divya)

2. Related Work

Sentiment analysis supports during the analysis of customer polarity on a particular product, information, or event. This task helps in understanding the attitude of the public, which helps in collecting reasonable information for future decisions on numerous comments. Sentiment analysis that was initially applied to political campaigns and news articles was then expanded to social media content. Recently this mission is used to capture feelings from Code-mix information available on social media.

Social media forum permits users to post content in informal settings. Also to enhance user experience, these forums allow the user to communicate their opinions in their native language or by switching between one or more languages according to their comfort. High resource language has formal settings that hold proper grammar rules. Earlier sentiment analysis model had grammatical rules and lexicons for extracting features from the input data. This rule-based feature extraction is complex and time-consuming.

To provide meaningful feature extraction and to make this a domain-independent task, enhancements are made in embeddings based on prominent features as an alternative to a rule-based system [11]. Those functions that convey the importance of the content are then fed into machine learning algorithms for performing multiclass classification. This classification model assigns various labels that help in understanding the sentiments of that data.

These systems do not work better for informal settings in the user-generated comments. Code-mixing and Code-switching alternate between two or more languages at document, phrase, sentence, token, lexeme and even at morpheme level [12]. In this enlarged usage of social media users, there arises a need for a model being trained with the Code-mixed language that functions on the user-generated comments [15]. This lead to the realization of the unavailability of a large dataset for Code-mixed language. This inspired the corpus collection of Code-mixed data from YouTube.

This annotated dataset is transformed to Term Frequency Inverse Document Frequency (TF/IDF) [22] representation and is applied to traditional Machine Learning algorithms for training. The traditional ML algorithms include Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Multinomial Naive Bayes (MNB) and, k-Nearest Neighbour (kNN). All these algorithms have performed well in classification tasks by classifying all the classes [13] [14], whereas SVM does not suit multiple class classification problems.

A multiple-task Kannada-English Code-Mixed dataset [17] for Sentiment Analysis and Offensive Language [18] detection has been collected, which consists of 7,671 comments that are annotated and are benchmarked using computational models. To promote multi-task learning for low-resourced languages, this dataset is used for training various classification models [16].

An enhanced technique in the processing of Code-Mixed language is by generating representations [9] of each sentence in the dataset. This representation gives the ability to learn the task-related features [10] from the input to facilitate classification. This representation is also generated by certain pre-trained models [19] [20] to understand the context of the input sentence. These features help in semantically classify the sentences based on the generated representation [21].

In this work, Sentiment Analysis on Code-Mixed Tamil language is performed by extracting the features in each sentence and classifying it based on the extracted features. The technique used for feature extraction and classification is explained in the subsequent topics.

3. Feature Extraction

Since the Machine learning model can't work with the raw data, some feature extraction techniques are applied to go on the raw data to convert it into vectors. Additionally, these models have been trained with certain training data to perform the task on the test data. Analyzing the similarity between the test data and the training data facilitates the classification process. To explore the similarity between the data, Term Frequency - Inverse Document Frequency (TF-IDF) [22] is applied. TF-IDF value is dependent on two factors.

- Term Frequency (TF) = No of times a specific word takes place in a document.
- Inverse Document Frequency (IDF) = Frequency of a term between the documents in the entire corpus.

The value of TF-IDF increases when a term often appears in the document and decreases when there are more documents in the entire corpus of that term. Thus, a high value is achieved when a term more often takes place in a document and the document appears less frequently in the corpus.

For each term, the TF-IDF score is computed and therefore the functional vector is framed for each sentence which is further fed as input for classification.

4. Classification Algorithm

The decision tree [7] is an effective and well-known classification algorithm. This algorithm generates a tree structure with the specified conditions to decide. Each node in the tree represents the state of an attribute and the result of this condition is represented using branches that connect each node. The labels are the judgments present in the leaf node. The decision tree may be error-free while handling classification problems with many labels but fewer samples. To overcome this disadvantage, a Random Decision Forest [8] classifier is developed.

Random Forest classifier builds a set of decision trees from the randomly selected subgroup of training data. The decisions taken by these trees are then collected and voting is carried out to make a final determination. This will be accomplished in the following steps.

- Choose random sample subgroups from the given dataset.
- Construct a decision tree for each subgroup sample and get a decision from each decision tree.
- The vote is done on each predicted decision.
- The decision with majority votes is made as to the final prediction.

This classifier is more accurate and vigorous in making decisions due to the numerous decision trees involved in the process. Even when the training is done with minimal samples, overfitting is ignored since the final decision is based on the average of numerous predictions that cancel the biases.

5. Proposed model

A system is proposed to perform multiclass classification on Code-mix data to detect the sentiment of that data. The input data must be classified into 5 groups for example in the positive, Negative, Mixed-feelings, Not-Tamil and unknown states. Initially, the input data must be pre-processed to remove symbols, special characters, hashtags and characters that do not hold any information. Preprocessed data is now represented as vectors using TF-IDF functional extraction technique. These vectors are assigned 5 labels using a random forestry classifier. Figure 1 illustrates the architecture of the proposed system.

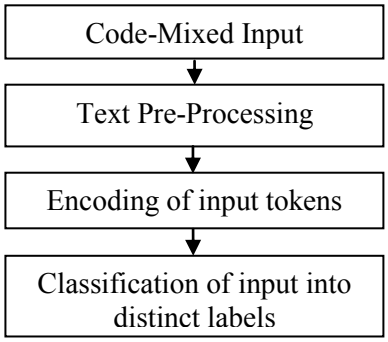

Figure 1: Architecture diagram for the proposed model

6. Performance Evaluation

The proposed model is applied to the Code-Mixed Tamil dataset collected from YouTube video comments. This dataset comprises 35,657 training sets, 3,962 validation sets and 4,403 test sets. The proposed model is trained using the training set and is evaluated using the validation set. The labels generated using the proposed model are assessed using the average weighted score for classification. The classification report for the proposed system is given below.

Table 1
Classification Report

	Precision	Recall	F1-Score	Support
Positive	0.56	0.50	0.53	2257
Negative	0.11	0.11	0.11	480
Mixed-feelings	0.13	0.04	0.07	438
Not-Tamil	0.08	0.29	0.13	176
Unknown state	0.16	0.18	0.17	611
Accuracy	-		0.35	3962
Macro avg	0.21	0.23	0.20	3962
Weighted avg	0.37	0.35	0.35	3962

The count of test data given for evaluating the system is mentioned as support. Weighted average F1 score is considered to assess the system developed for assigning labels. This is calculated as the average of precision and recall.

7. Conclusion and Future work

Identification of sentiments in Code-mix Tamil data is done using a machine learning classifier and the evaluation of the proposed system is accomplished. Applying profound learning techniques will further enhance the learning of the model and will enhance classification performance.

8. Acknowledgements

We sincerely thank the management of SSN Institutions for the infrastructure and lab facilities to carry out this research work.

9. References

[1] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." In Proceedings of human language technology conference and conference on empirical methods in natural language processing, pp. 347-354. 2005.

[2] Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J. Passonneau. "Sentiment analysis of Twitter data." In Proceedings of the workshop on language in social media (LSM 2011), pp. 30-38. 2011.

[3] Thavareesan, Sajeetha, and Sinnathamby 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), pp. 320-325. IEEE, 2019.

[4] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." arXiv preprint cs/0409058 (2004).

[5] Pratapa, Adithya, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. "Language modeling for code-mixing: The role of linguistic theory based synthetic

data." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1543-1553. 2018.
[6] Barman, Utsab, Amitava Das, Joachim Wagner, and Jennifer Foster. "Code mixing: A challenge for language identification in the language of social media." In Proceedings of the first workshop on computational approaches to code-switching, pp. 13-23. 2014.
[7] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21, no. 3 (1991): 660-674.
[8] Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.
[9] Banerjee, Shubhanker, Bharathi Raja Chakravarthi, and John P. McCrae. "Comparison of pre-trained embeddings to identify hate speech in Indian code-mixed text." In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 21-25. IEEE, 2020.
[10] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Shubhanker Banerjee, Richard Saldanha, John Philip McCrae, Parameswari Krishnamurthy, and Melvin Johnson. "Findings of the Shared Task on Machine Translation in Dravidian languages." In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 119-125. 2021.
[11] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, R. L. Hariharan, John Philip McCrae, and Elizabeth 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, pp. 133-145. 2021.
[12] Hande, A deep, Siddhanth U. Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. "Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced Dravidian languages." arXiv preprint arXiv:2108.03867 (2021).
[13] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. "Overview of the track on sentiment analysis for Dravidian languages in code-mixed text." In Forum for Information Retrieval Evaluation, pp. 21-24. 2020.
[14] Chakravarthi, Bharathi Raja, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. "A sentiment analysis dataset for code-mixed Malayalam-English." arXiv preprint arXiv:2006.00210 (2020).
[15] Chakravarthi, Bharathi Raja, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. "Corpus creation for sentiment analysis in code-mixed Tamil-English text." arXiv preprint arXiv:2006.00206 (2020).
[16] Mandl, Thomas, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. "Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German." In Forum for Information Retrieval Evaluation, pp. 29-32. 2020..
[17] Hande, A deep, Ruba Priyadharshini, and Bharathi Raja 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, pp. 54-63. 2020.
[18] Mandl, Thomas, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. "Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german." In Forum for Information Retrieval Evaluation, pp. 29-32. 2020.
[19] Ghanghor, Nikhil, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. "IIITK@ DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada." In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 222-229. 2021.
[20] Banerjee, Shubhanker, Arun Jayapal, and Sajeetha Thavareesan. "NUIG-Shubhanker@ Dravidian-CodeMix-FIRE2020: Sentiment Analysis of Code-Mixed Dravidian text using XLNet." arXiv preprint arXiv:2010.07773 (2020).

[21] Suryawanshi, Shardul, and Bharathi Raja 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, pp. 126-132. 2021.
[22] Aizawa, Akiko. "An information-theoretic perspective of tf–idf measures." Information Processing & Management 39, no. 1 (2003): 45-65.