Parameswari_faith_nagaraju@Dravidian-CodeMix-FIRE: A machine-learning approach using n-grams in sentiment analysis for code-mixed texts: A case study in Tamil and Malayalam

Parameswari Krishnamurthy[a], Faith Varghese[b] and Nagaraju Vuppala[c]

[a]*University of Hyderabad, India*
[b]*Independent Researcher, India*
[c]*eBhashasetu Language Services Pvt Ltd, India*

Abstract

Sentiment analysis is a fast growing research positioned to uncover the underlying meaning of a text by categorizing it into different levels. This paper is an attempt to decode the deeply entangled code-mixed Malayalam and Tamil datasets and classify its interlined meaning at five various levels. Along with the corpus creation, [1] propose a five-level classification for Malayalam and Tamil code-mixed datasets. In this paper, we follow the five-level annotated datasets and aim to solve the classification problem by implementing unigram and bigram knowledge with a Multinomial Naive Bayes model. Our model scores an F1-score of 0.55 for Tamil and 0.48 for Malayalam.

Keywords

Sentiment Analysis, Code-mixed texts, n-gram, a Multinomial Naive Bayes model, Tamil, Malayalam

1. Introduction

Sentiments are constructed using elements to express positive or negative sentiments and sentiment analysis thus detects the opinion of the sentence/document and classifies it into positive, negative or neutral. [2, 3] try to address four different problems predominating in this research community, namely, subjectivity classification, word sentiment classification, document sentiment classification, and opinion extraction. The automated process of discerning or monitoring the opinions about a given subject, not only assists us in training the machine to associate certain inputs with the corresponding outputs but also spots the keywords to assess the stance of the consumer, to scan its polarity. The proliferation of commercial applications has been one of the major reasons for the flourishing of sentiment analysis in the industrial field as well. This provides a strong motivation for research on Tamil and Malayalam code-mixed data in sentiment analysis and offers many challenging research problems, which would have been tough to address, otherwise. Presently, sentiment analysis is the cynosure of social media research. Tamil and Malayalam, a widely used language in social media in different domains

FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India

EMAIL: pksh@uohyd.ac.in (P. Krishnamurthy); faith.varghese@gmail.com (F. Varghese); vpp.nagaraju1234@gmail.com (N. Vuppala)

CEUR Workshop Proceedings (CEUR-WS.org)

needs a robust sentiment analysis. Starting from the assessment of marketing the success of an ad campaign or new product launch, to determining the versions of a product or service that are popular, and even identifying the demographics of people's likes and dislikes particularly, is a contribution much needed in these languages. Hate speech detection is another inevitable tantamount achievement of this domain. In this paper, an n-gram knowledge-base trained with a Multinomial Naive Bayes model is employed to analyse code-mixed texts in sentiment analysis for Tamil and Malayalam.

Tamil and Malayalam, the major Dravidian languages, are agglutinative i.e. word may contain multiple morphemes attached to the stem with distinct morpheme boundaries to form a multimorphemic word [4, 5, 6]. Understanding certain linguistic features which are encoded as inflection and derivation is crucial in sentiment classification. According to a KPMG report (2017)[1], Tamil and Malayalam users of the internet are 42% and 27% respectively. Social media platforms have millions of Indian language users and it is evident that due to bilingualism and multilingualism in India, the code-mixed use of language is a common factor.

The majority among the vast group of social media users with bilingual or multilingual proficiency prefer to adopt code-mixing as it conveys the concept in the most simplest and acceptable fashion [7, 8]. It is a normal tendency that the internet users opt for Roman transliteration over the native scripts when they comment in social media. In some cases, users mix more than a language in their usage resulting in code-mixing. Similar to this concept another form we witness is the usage of combination of more than a script in expressing the idea. The conventional approach of feature based classification may not assist us in yielding an optimum prediction as the code-mixed language structure and spelling cannot be predefined. Implementation of sentiment analysis in Indian languages is so recent that its emergence is noted by the end of the first decade of the 21st century and however, extensive research has not been reported in Tamil or Malayalam code-mixed corpora.

2. Brief Survey

A few efforts on extracting sentiments in Tamil and Malayalam are reported here.

- The research by [9] focus on domain-specific sentence-level mood extraction from Malayalam text using two methods of sentiment analysis viz., machine learning method and semantic orientation method. The task is carried out using a semantic orientation method using *pointwise mutual information retrieval* algorithm.
- the research article "SentiMa - Sentiment Extraction for Malayalam" by [10] propounds a rule-based approach for opinion analysis from Malayalam movie reviews. The rule-based approach that has been suggested by the researcher for extracting the opinion analysis in Malayalam is the Negation-Rule that has claimed to have achieved 85% accuracy.
- Authors in [11] employed sentiment analysis on the tweets in three languages, namely, Hindi, Bengali, and Tamil for datasets collected from twitter over a period of three months. The purpose of this shared task was to classify the collected tweets into positive, negative and neutral polarity. From the submissions of six teams, maximum accuracy attained

[1] https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

were 43.2 %, 55.67 %, and 39.28 % for Bengali, Hindi and Tamil respectively and to achieve this accuracy the teams had employed supervised classification algorithms such as Naïve Bayes, Multinomial Naïve Bayes, Support Vector Machines and Decision Tree.

- Authors in [12] propose a N-gram model for opinion classification of Tamil tweets. 7418 Tamil unicode tweets were manually annotated by various domain experts for this study and it aimed at three level classification. 61.29% is the overall accuracy estimated by the model in this approach.
- The lexicon-based method is used by [13] uses which is also effectively practiced [14, 15, 16] for Malayalam as it is crucial to identify functional features along with the lexical categories, as certain linguistic features are encoded as inflection and derivation . For this study, 87347 Malayalam unicode sentences with political content were selected from news web resources in the period between 1st August 2018 and 30th September 2018 and the feature-based classification assisted to achieve a f-score of 0.9290.
- In order to address the decision problem of code-mixed Tamil and Malayalam texts in sentiment analysis, [17] [1], [18], [19] built a corpus of Tamil-English and Malayalam-English by manually annotating 15,744 and 6,738 YouTube comments respectively. They employ five-level classification for annotation: Positive, Negative, Neutral, not-Malayalam or not-Tamil and Mixed feelings. The annotation derived at an agreement of Krippendorff's alpha 0.6 and 0.89 for Tamil and Malayalam respectively.

3. Trained Dataset: An Analysis

Dravidian-CodeMix Forum for Information Retrieval Evaluation (FIRE) 2020 releases the datasets of Tamil with 11335 sentences and Malayalam 4851 sentences. An overview of the categories present in the dataset is given below:

Category	Tamil	Category Percentage	Malayalam	Category Percentage
Positive	7627	67.29	2022	41.68
Negative	1448	12.77	549	11.32
Mixed_feelings	1283	11.32	289	5.95
Unknown_state	609	5.37	1344	27.71
Not-Tamil/Malayalam	368	3.25	647	13.34

Table 1: Tamil and Malayalam train dataset

As seen in Table 1, the dataset covering the Positive category is higher than any other categories both in Tamil and Malayalam. The initial hypothesis is that the class-imbalance in the training data may adversely affect the model accuracy. Furthermore, it is to be also noted that building a balanced corpus from diverse YouTube comments is a cumbersome task. Hence, we adopt n-gram modelling to examine whether imbalanced data with code-mix for sentiment analysis can be handled.

4. Technique used

The algorithm used in the present research in sentiment analysis for code-mixed data is described in the following stages.

1. Preprocessing
2. N-Gram modelling
3. Conversion into weighted features and machine learning

4.1. Preprocessing

The input data is preprocessed which involve dataset cleaning, removing stop words, punctuations, numbers and non-unicode characters. As the dataset is given in Roman script, the text is converted into lowercase.

4.2. N-gram modelling

From the pre-processed trained dataset, bigrams and unigrams are identified with highest probabilistic sentiment category. The unigram and bigram database of Tamil are 18,195 and 58,129 respectively. Similarly in Malayalam 12,356 and 28,291 respectively. Each sentence is converted into a unigram and bigram model based on the maximal match.

4.3. Conversion into weighted features and machine learning

There are many techniques in machine learning that can be used to categorize data. The present task used Term Frequency- Inverse Document Frequency (TF-IDF) approach [20]. TF-IDF is a numerical statistic that shows the relevance or importance of a word/keyword to a document in a collection of corpus. After experimenting with the Bag of words approach, we observed that TF-IDF gives better results. We have used it as an initial step is to convert the data(training and testing) into numerals. These numerals are feature vectors, i.e each vector has its own importance. Since we are dealing with the classification of texts into positive, negative, unknown_state, mixed_feelings, not Malayalam, not Tamil, Multinomial Naive Bayes (NB) Algorithm [20] is used to classify text/comments into categories. This model works efficiently when there are multiple categories with the combination of TF-IDF features. Multinomial NB model and trains each sentence with its given category as shown in the flowchart-1 [2]

5. Results and Discussion

The training dataset of Tamil includes 3149 sentences and Malayalam consists of 1348 sentences. An overview of the categories present in the testing dataset is given below:

[2] The code can be accessed here: https://github.com/nagaraju291990/sentimentAnalysis

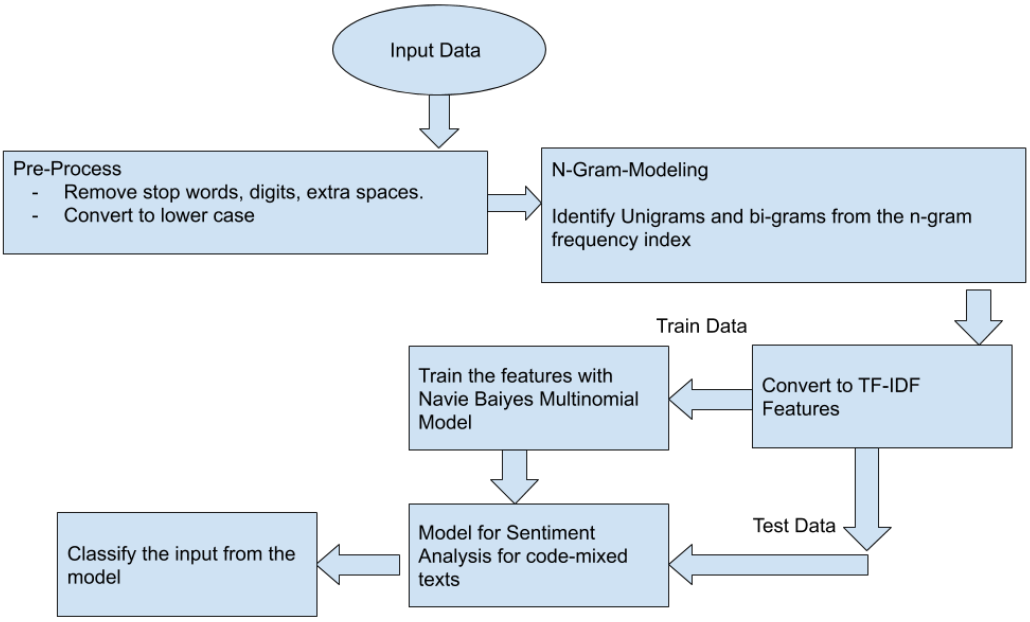
Input Data
Pre-Process
- Remove stop words, digits, extra spaces.
- Convert to lower case
N-Gram-Modeling
Identify Unigrams and bi-grams from the n-gram frequency index
Train Data
Convert to TF-IDF Features
Train the features with Navie Baiyes Multinomial Model
Test Data
Model for Sentiment Analysis for code-mixed texts
Classify the input from the model

Figure 1: Flowchart

Category	Tamil	Category Percentage	Malayalam	Category Percentage
Positive	2075	65.89	565	41.91
Negative	424	13.47	138	10.24
Mixed_feelings	377	11.97	70	5.19
Unknown_state	173	5.49	398	29.53
Not-Tamil/Malayalam	100	3.18	177	13.13

Table 2: Tamil and Malayalam test dataset

The table 3 provides the results of the dataset given in table- 2.

Language	Precision	Recall	F-score
Tamil	0.55	0.66	0.55
Malayalam	0.53	0.51	0.48

Table 3: Results

In the case of identifying positive category, our model performs better due to the datasize.

Whereas in identifying other categories, the failure occurred. Similarly, as noticed there are cases in which the annotated data has some errors too:

For instance, in Malayalam as seen in sentence (1), it is actually annotated as Not-Malayalam, however it expresses Positive. Similarly, sentence (2) is tagged as Negative, however it denotes Positive.

(1) ml_sen_468:
`Proud to be a vypinkkari....` Not-malayalam
'Proud to be a Vypin lady'

(2) ml_sen_1319:
`Ithinu apurath oru trailer illa` Negative
'There is no trailer beyond this'

Similarly, for instance in Tamil sentence (3) is tagged as positive, but it is not-Tamil and sentence (4) expresses negative feeling, however tagged as Positive.

(3) ta_sent_22:
`I am simbu fans like dhanush acting` Positive
'I am the fan of Simbu and like Dhanush acting'

(4) ta_sent_29:
`Inum na sagurathu kulla intha mathiri yethana kodumaiya paka poranoo therilayee` Positive
''I dont know how many inhuman act I will witness like this before I die"

Another inevitable reason is the n-gram model and its size. Higher the model learns the context the better the prediction would be; and the major factor that resulted in imprecise labelling is the shortfall of the n-gram training in the given category. Also there are inconsistencies in the data regarding the distribution of categories. For example positive categorised data amounts to 65% while the remaining categories each fall under 15% or much less percentage.

6. Conclusion

The possibility of sentiment analysis in Tamil and Malayalam code-mixed texts by applying the combination of unigram and bigram approach is discussed in this paper. The model accuracy can be improved with the increasing in the database of the unigrams and bigrams. On the other hand Multinomial NB machine learning model works well when we need to train the dataset with more than two sentiments while other models work well with Boolean sentiment classification. To further improve the model, spelling normalization of code-mixed data and integrating linguistic methods would be adopted in the future study. Although this requires a lot of data and accurately human classified sentences to improve the accuracy of the machine learning model, the distribution of sentiments equally in the train dataset is another important aspect for machine learning model not to be biased to a particular sentiment.

References

[1] 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.

[2] B. Liu, Sentiment analysis and opinion mining (series synthesis lectures on human language technologies). vol. 16, San Mateo, CA, USA: Morgan (2012).

[3] H. Tang, S. Tan, X. Cheng, A survey on sentiment detection of reviews, Expert Systems with Applications 36 (2009) 10760–10773.

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

[5] 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), volume 70 of *OpenAccess Series in Informatics (OASIcs)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019, pp. 6:1–6:14. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10370. doi:10.4230/OASIcs.LDK.2019.6.

[6] B. R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. S, M. Arcan, M. Zarrouk, J. P. McCrae, Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 56–63. URL: https://www.aclweb.org/anthology/W19-6809.

[7] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, J. P. McCrae, A survey of current datasets for code-switching research, in: 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS), India, 2020.

[8] R. Priyadharshini, B. R. Chakravarthi, M. Vegupatti, J. P. McCrae, Named entity recognition for code-mixed Indian corpus using meta embedding, in: 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS), India, 2020.

[9] N. Mohandas, J. P. Nair, V. Govindaru, Domain specific sentence level mood extraction from malayalam text, in: 2012 International Conference on Advances in Computing and Communications, IEEE, 2012, pp. 78–81.

[10] D. S. Nair, J. P. Jayan, R. Rajeev, E. Sherly, Sentiment analysis of malayalam film review using machine learning techniques, in: 2015 international conference on advances in computing, communications and informatics (ICACCI), IEEE, 2015, pp. 2381–2384.

[11] B. G. Patra, D. Das, A. Das, R. Prasath, Shared task on sentiment analysis in indian languages (sail) tweets- an overview, in: International Conference on Mining Intelligence and Knowledge Exploration, Springer, 2015, pp. 650–655.

[12] N. Ravishankar, R. Shriram, Grammar rule-based sentiment categorisation model for classification of tamil tweets, International Journal of Intelligent Systems Technologies and Applications 17 (2018) 89–97.

[13] F. T. Varghese, A computational implementation of opinion analysis: a case study of

malayalam political texts on social media, Unpublished Mphil dissertation: University of Hyderabad (2018).

[14] P. D. Turney, M. L. Littman, Unsupervised learning of semantic orientation from a hundred-billion-word corpus, arXiv preprint cs/0212012 (2002).

[15] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2011) 267–307.

[16] P. Chesley, B. Vincent, L. Xu, R. K. Srihari, Using verbs and adjectives to automatically classify blog sentiment, Training 580 (2006) 233.

[17] 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.

[18] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Davidian 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.

[19] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth McCrae, Overview of the track on Sentiment Analysis for Davidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.