CIA_NITT@Dravidian-CodeMix-FIRE2020: Malayalam-English Code Mixed Sentiment Analysis Using Sentence BERT And Sentiment Features

Yandrapati Prakash Babu, Rajagopal Eswari and K Nimmi

Department of Computer Applications, National Institute of Technology, Trichy, India.

Abstract

Code mixing is the mixing of language while writing text. The biggest problem in Malayalam-English code-mixing is that people switch between languages (e.g. Malayalam and English) and use English phonetic typing instead of writing Malayalam words using Dravidian scripts. Traditional NLP models are trained on extensive monolingual tools (e.g. Malayalam or English); code-mixing is challenging since they cannot handle data code-mixing. DravidianCodeMix FIRE 2020 is to classify comments into positive, negative, unknown_state, mixed_feelings and not-malayalam categories based on message-level polarity using Malayalam - English dataset. The classification model used for this challenging task is sentence-level BERT. Achieved an F1-score of 0.71 (ranked 4th) for the Malayalam-English (Manglish) comments dataset submitted under the username: CIA_NITT.

Keywords

BERT, Code Mixed, Manglish, Sentiment Analysis

1. Introduction

In multilingual countries like India which has 22 official languages, code-mixing is often noted. Dravidian languages are spoken by 19.64 percent of Indians; not enough research has been done to address these languages' sentiment analysis [1, 2, 3]. The DravidianCodeMix FIRE 2020 task's objective is to define the code-mixed dataset's sentiment polarity obtained from social media comments/posts in the Dravidian Language Malayalam-English (Manglish). The Manglish language comprises words from Malayalam as well as English in the same sentence written using English alphabets. The dataset provided is YouTube comments written in Manglish[4]. This initiative aims to encourage research that will show how opinion is expressed on social media in code-mixed scenarios. The label column should have only the labels 'not-Malayalam', 'unknown_state', 'Positive', 'Mixed_feelings', 'Negative'.

According to Myers et al. Code Mixing (CM) [5] refers to integrating an utterance of another language of linguistic units such as sentences, terms and morphemes of one language. Code-mixed languages can contain words from multiple languages. Here the main focus is only on Code-mixed bilingual language (Malayalam-English) [6, 7]. Code-mixing is very common in a multilingual culture, and code-mixed texts are often written in non-native scripts [8, 9]. Code

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

EMAIL: prakash.babu23@gmail.com (Y. Prakash Babu); eswari@nitt.edu (R. Eswari); menimmicse@gmail.com (K. Nimmi)

mixing which is done here is Malayalam sentences written using English alphabets [10, 11]. Code-mixing is the word-level alternation of languages that often occurs by combining words from one language with another language's rules, according to Solorio et al. [12]. Language mixing when writing text, also known code-mixing. Natural language processing (NLP) is a modern technology that provides computers with knowledge in order to understand the languages humans speak. NLP involves syntax analysis (grammatical rules) and semantic analysis (meaning). Sentiment analysis is a technique of classification that collectively provides feelings relevant to a subject. Sentiment analysis may be carried out at sentence level, document level, aspect level and phrase level. Sentiment Analysis is a concept which is commonly used to describe a human's emotional states. We did not find any research on Manglish Corpora in sentiment analysis to the best of our knowledge. The task organizers created the datasets for Malayalam-English and Tamil-English, They clearly explained how they gathered and labeled the comments in the datasets [4, 13]. This paper suggests an analysis of sentiments using sentence BERT[14] for comments in Malayalam-Language (Manglish).

2. Related Research work

Related studies on code-mixing computing models are less since there is a scarcity of conventional text corpora. Joshi et al. carried out pioneering work in the sentiment analysis of Hindi language, in which the authors developed a three-step fallback model [15] based on machine translation, lexicons of sentiment and classification. The framework performed best with unigram features. Vyas et al has been using Parts of Speech (POS) [16] for Hindi sentences embedded with English words and found that Hindi language recognition and transliteration are two significant challenges that affect the accuracy of the POS tagging. Prabhu et al. proposed an LSTM (Sub word-LSTM) [17] that function well on extremely noisy text and text containing misspellings they got a 4-5% greater accuracy than traditional approaches. Code-mixed noisy content cleaning at word-level based on orthographic information [18]. Text classification using deep learning models to identify the Manglish and Tanglish sentiments [19, 20]. Kumar et al. used an ensemble model [21] for mixed code tweet analysis and achieved an F1 score of 0.70 for Hindi-English (Hinglish) and 0.725 for Spanish-English (Spanglish) datasets.

For code-mixed social media text analysis, Singh et al. used cross-lingual embeddings [22], which is unsupervised, and obtained an F1 score of 0.6355. Sharma et al. developed a shallow Hindi-English code-mixed social media text parser[23]; this shallow parser model was modelled as three separate sequence labeling problems. Advani et al. built a classifier that can use [24] hand-engineered lexical, sentiment, and metadata features to distinguish between "positive", "neutral" and "negative" feelings. Thomas et al. performed Sentimental Study of Transliterated Text using RNN-LSTM [5] technique to extract transliterated text sentiments. Das et al.[25] suggested a computational technique to produce a SentiWordNet equivalent for Bengali from publicly accessible English Sentiment lexicons and bilingual English-Bengali dictionary. Bhargava et al. used the Language Recognition [26] and Sentiment Mining Method with the combination of four languages. A novel hybrid architecture of deep learning [27] that is highly successful in analysing emotions in resource poor languages based on four Hindi datasets covering various domains was proposed by Akhtar et al. A rule-based, n-gram, multivariate feature selection system was developed by Abbasi et al .[28].

3. Methdology

3.1. Data and Pre-Processing

The task organizers provide the dataset collected from YouTube[1]. The given dataset has multiple polarities like positive, negative, not-malayalam, unknown_state and mixed_feelings. The comments in the dataset are noisy, to clean the comments, we applied pre-processing steps like convert comment into lower case, removing the special characters, replace the emojis with related word, removing the repeating characters which are appear in the word more than two times. The dataset has the class imbalance problem, class labels not-malayalam, unknown_state and mixed_feelings have less sample compared to positive and negative class labels. The statistics of the each class is tabulated in the Table 1.

Labels	**Training**	**Validation**	**Test**
Positive	2022	224	565
Negative	1344	161	398
Not-malayalam	647	60	177
Unknown_state	549	51	138
Mixed_feelings	289	44	70

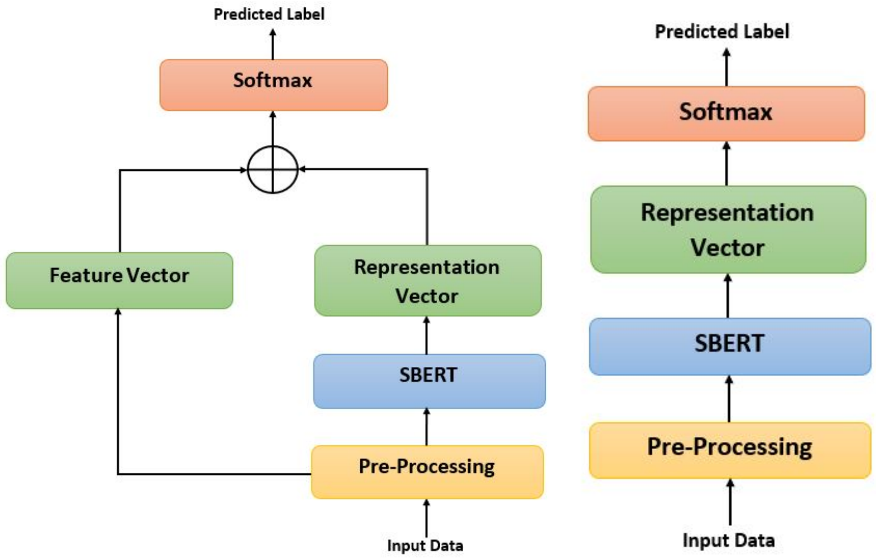
Table 1: Statistics of Training, Validation and Test sets

3.2. Model Description

Model-1 is based on Sentence BERT (SBERT)[14] and Manglish features. In this work 252 Manglish features(Manglish sentiment words) are manually gathered from the YouTube comments. The whole SBERT model is finely tuned using the training dataset. The original comment is added with the [CLS] and [SEP] special tokens and then word-piece tokenizer is applied for tokenization. The fusion of each token is accomplished by the summation of the word-piece's embedding, position, and segment. A series of 12 transformer encoder layers is applied to obtain the final hidden state vectors on these token embeddings. Following , we treat $V_1 = SBERT(comment) \in R^{768}$ the final hidden vector of [CLS] token as the representation of comment, and for the feature vector representation $V_2 = R^N$ where N = number of features, one vector is generated for every comment with the length of 252 having 0 and 1 (if Manglish sentiment word is appeared in the comment 1 is appended to the vector otherwise 0 is appended). Then, V_1 and V_2 is concatenated $V = V_1 \oplus V_2$ where $\oplus$ represents concatenation, passed through fully connected softmax layer to get the required label $\hat{y} \in R^C$, where C=number of classes. Figure 1(a) shows the methodology of Model-1.

$$\hat{y} = Softmax(W^T V + b) \tag{1}$$

[1] https://dravidian-codemix.github.io/2020/datasets.html



(a) SBERT with Feature Vector (b) SBERT with CBL

Figure 1: Overview of the Models used in this proposed work

Model-2 is also based on SBERT but instead of cross entropy loss, Class Balanced Loss (CBL) [29] is used to handle the imbalance problem in the dataset. Here $\alpha_t = (1-\beta)/(1-\beta^{n_y})$, α_t is the balanced focal loss factor, C denotes total number of classes, y denotes labels from 1,2,..,c, p_i denotes probabilities varies from 0 and 1, β and γ are the hyper parameters. Figure 1(b) shows the methodology of Model-2.

$$CBL = -\alpha_t \sum_{i=1}^{C} (1-p_i^t)^\gamma \log(p_i^t) \tag{2}$$

4. Implementation Details

In the Model-1 SBERT with Feature Vector is used and to train this model. SBERT Hyper parameters are set to epochs=3, batch size = 32, learning rate = 3e-5 and dropout=0.4. In Model-2 SBERT with Class Balanced Loss is used, for this we set additional hyperparameters as β=0.9999, γ = 2.0. Two models implemented using the transformers library in PyTorch [30]. The implementation code is available in the github [2].

[2] https://github.com/prakashbabuy/manglish/

5. Results

Label	precision	Recall	F1-score
Mixed_feelings	0.25	0.30	0.27
Negative	0.67	0.25	0.37
Positive	0.82	0.58	0.67
not-malayalam	0.73	0.70	0.71
unknown_state	0.50	0.81	0.62

Table 2: Precision, Recall and F1-score of SBERT on evaluation set

Label	Precision	Recall	F1-score
Mixed_feelings	0.39	0.46	0.42
Negative	0.71	0.43	0.54
Positive	0.85	0.69	0.76
not-malayalam	0.72	0.85	0.78
unknown_state	0.62	0.79	0.69

Table 3: Precision, Recall and F1-score of SBERT with feature vector on evaluation set

Label	precision	Recall	F1-score
Mixed_feelings	0.42	0.39	0.40
Negative	0.65	0.48	0.55
Positive	0.77	0.80	0.79
not-malayalam	0.84	0.73	0.78
unknown_state	0.63	0.71	0.67

Table 4: Precision, Recall and F1-score of SBERT with CBL on evaluation set

Model	Precision	Recall	F1-Score
SBERT	67	61	61
SBERT With Feature Vector	72.4	70.32	70.29
SBERT with CBL	71	71	71

Table 5: Precision, Recall and F1-score of proposed models on evaluation set

To classify Code-Mixed Malayalam and English comments, we experimented with three models namely SBERT, SBERT with Feature Vector and SBERT with CBL. The class wise SBERT performance is shown in Table 2. Due to class imbalance, SBERT model performance is limited. The performances of SBERT with Feature Vector and SBERT with CBL models are reported in the Tables 3 and 4. SBERT with Feature Vector model achieved F1-score of 70.29% and SBERT with CBL model achieved F1-score of 71%. From the Table 5, it is clear that SBERT with

CBL model achieved slightly better results compared to SBERT with Feature Vector. Both our proposed models achieved better performance compared to SBERT.

6. Conclusion

This paper presents two models to classify Code-Mixed Malayalam and English comments models based on Sentence BERT. This task is treated as multiclass text classification problem. The model based on SBERT with CBL achieved F1-score of 71%. In the future we will improve the model to find the sarcastic Manglish comments.

References

[1] B. R. Chakravarthi, M. Arcan, J. P. McCrae, Improving Wordnets for Under-Resourced Languages Using Machine Translation, in: Proceedings of the 9th Global WordNet Conference, The Global WordNet Conference 2018 Committee, 2018. URL: http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_16.

[2] 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.

[3] B. R. Chakravarthi, M. Arcan, J. P. McCrae, WordNet gloss translation for under-resourced languages using multilingual neural machine translation, in: Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 1–7. URL: https://www.aclweb.org/anthology/W19-7101.

[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://www.aclweb.org/anthology/2020.sltu-1.25.

[5] C. Myers-Scotton, Duelling languages: Grammatical structure in codeswitching, Oxford University Press, 1997.

[6] 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 and Communication Systems (ICACCS), 2020.

[7] 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 and Communication Systems (ICACCS), 2020.

[8] 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.

[9] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, P. Buitelaar, A dataset for troll classification of Tamil memes, in: Proceedings of the 5th Workshop on Indian Language Data Resource and Evaluation (WILDRE-5), European Language Resources Association (ELRA), Marseille, France, 2020.

[10] B. R. Chakravarthi, P. Rani, M. Arcan, J. P. McCrae, A survey of orthographic information in machine translation, arXiv e-prints (2020) arXiv–2008.

[11] 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 for Hindi-English code-mixed data, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020.

[12] T. Solorio, Y. Liu, Learning to predict code-switching points, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 973–981.

[13] 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.

[14] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://www.aclweb.org/anthology/D19-1410. doi:10.18653/v1/D19-1410.

[15] A. Joshi, A. Balamurali, P. Bhattacharyya, et al., A fall-back strategy for sentiment analysis in hindi: a case study, Proceedings of the 8th ICON (2010).

[16] Y. Vyas, S. Gella, J. Sharma, K. Bali, M. Choudhury, POS tagging of English-Hindi code-mixed social media content, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 974–979.

[17] A. Prabhu, A. Joshi, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text, arXiv preprint arXiv:1611.00472 (2016).

[18] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020. URL: http://hdl.handle.net/10379/16100.

[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 Dravidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[20] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly,

Elizabeth McCrae, Overview of the track on Sentiment Analysis for Dravidian 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.

[21] A. Kumar, H. Agarwal, K. Bansal, A. Modi, BAKSA at SemEval-2020 task 9: Bolstering CNN with self-attention for sentiment analysis of code mixed text, arXiv preprint arXiv:2007.10819 (2020).

[22] 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.

[23] 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).

[24] 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).

[25] A. Das, S. Bandyopadhyay, Subjectivity detection in English and Bengali: A CRF-based approach, Proceeding of ICON (2009).

[26] R. Bhargava, Y. Sharma, S. Sharma, Sentiment analysis for mixed script Indic sentences, in: 2016 International conference on advances in computing, communications and informatics (ICACCI), IEEE, 2016, pp. 524–529.

[27] M. S. Akhtar, A. Kumar, A. Ekbal, P. Bhattacharyya, A hybrid deep learning architecture for sentiment analysis, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 482–493.

[28] A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, IEEE Transactions on Knowledge and Data Engineering 23 (2010) 447–462.

[29] Y. Cui, M. Jia, T. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9260–9269.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).