

A Few Shot Learning to Detect Sarcasm in Tamil and Malayalam Code Mixed Data

Shanmitha Thirumoorthy¹, Manavh N R¹, Durairaj Thenmozhi² and Ratnavel Rajalakshmi¹

¹Vellore Institute of Technology, Chennai

²Sri Sivasubramaniya Nadar College of Engineering, Chennai

Abstract

Sarcasm poses significant challenges in sentiment analysis. With the intended meaning differing from the literal one, it subtly conveys a viewpoint. On social media communications, which are frequently code mixed for Dravidian languages, there is an increasing demand for sarcasm identification to detect the correct sentiments. Sarcasm identification shared task at FIRE 2023 aims to detect sarcasm in Tamil-English and Malayalam-English code mixed data collected from Youtube comments. A few shot learning approach is employed to identify whether the comments are sarcastic in Dravidian code mixed languages. 2-way-20-shot variation with Paraphrase-MiniLM-L3-v2 embeddings and logistic regression as a classifier gives F1 scores of 0.68 and 0.57 for Tamil-English and Malayalam-English data sets respectively. Our team Hydrangea secured sixth position in the leader board for both data sets.

Keywords

Sarcasm Identification, Few Shot Learning, Deep Learning, Sentiment Analysis, Text Analytics

1. Introduction

Sarcasm is the use of words that have a different meaning than what you truly mean to express, often to offend or irritate someone or humorously criticise something. Detecting sarcasm is very much important in sentiment analysis. In sentiment analysis, the sentiment categories are very clearly defined, however the borders of sarcasm are not that well defined. Thus, presence of sarcasm in text considerably affects the performance of sentiment analysis and also in other applications like homophobia detection [1] and hope speech identification [2]. Identifying sarcasm is a challenging task for Dravidian languages in specific.

Sarcasm detection is a popular research field and several research works have been reported in English [3][4][5] and in European languages [6]. A few methodologies are reported in literature on detecting sarcasm in Hindi [7][8] and Hindi-English code mixed data [9][10]. However, sarcasm detection in Dravidian languages are in a premature stage. Sarcasm_Identification_DraavidianCodeMix@FIRE-2023 [11][1][2] focuses on detecting sarcasm from Tamil-English and Malayalam-English code mixed data.

Forum for Information Retrieval Evaluation, December 15-18, 2023, India

✉ shanmitha.t2023@vitstudent.ac.in (S. Thirumoorthy); manavh.nr2023@vitstudent.ac.in (M.N. R);

theni_d@ssn.edu.in (D. Thenmozhi); rajalakshmi.r@vit.ac.in (R. Rajalakshmi)

🆔 (0000-0003-0681-6628 (D. Thenmozhi)



© 2023 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)

Several methodologies such as traditional learning with word embeddings, deep neural networks and transformers were used by researchers to detect sarcasm. Few shot learning is very popular in image analysis [12][13][14][15]. However, a very few works have been reported on text applications such as short text classification [16], sentiment analysis [17] and named entity recognition [18]. In this paper, few shot learning is employed to detect sarcasm in Tamil-English and Malayalam-English code mixed data. A few shot learning is a branch of machine learning and deep learning, which is to teach AI models how to learn with just a little amount of labelled training data. Its objective is to provide models the ability to generalise additional, unforeseen data samples based on a limited number of samples we provide them throughout the training phase. The other models namely XLM RoBERTa, mBERT and BERT were also used to identify the sarcasm and to compare with the proposed approach.

2. Related Work

Eke et al. [19] used a combination of the BERT model, and traditional machine learning to present a context-based feature approach for sarcasm detection. They used the Internet Argument Corpus, version two (IAC-v2) dataset for evaluation. They have employed three models. The first model builds GloVe embeddings with bidirectional long short term memory. The second model is built on BERT model with a pre-trained Bidirectional Encoder representation. The third model is an ensemble model with BERT and GloVe embedding features along with traditional machine learning model. Onan [20] states that Topic-enriched word embedding scheme's predictive performance on sarcasm identification has been improved with that of traditional word-embedding techniques namely word2vec, fastText, and GloVe. They used the other standard lexical, pragmatic, implicit, and explicit incongruity-based features in addition to word-embedding-based features for detecting sarcasm. They have evaluated on Twitter messages.

Parveen et al. [21] used a CNN model incorporating both implicit and explicit representations of brief text for the purpose of classifying sarcasm. They collected data from Twitter and Amazon to evaluate their approach. Pandey and Singh [22] established a model made up of Long Short Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT-LSTM). The code mixed dataset is embedded using a pre-trained BERT model. A single-layer LSTM network that used these embedding vectors was used to determine if a statement was sarcastic or not.

Kalaivani and Thenmozhi [5] used deep learning approach (LSTM -RNN) and BERT alongside traditional approaches to identify sarcasm. Using these approaches, they built the model, identified and categorized the response quantity required for the detection of sarcasm on the two forums of Twitter and Reddit. Pawar and Bhingarkara [8] proposed a pattern-based approach with four sets of features that include a lot of details about sarcasm to detect sarcastic Tweets.

3. Dataset Description

Sarcasm_Identification_Draavidian-CodeMix@FIRE-2023 [11] shared task provides data sets namely training, development and test data for two languages namely Tamil-English and

Table 1
Data Distribution

Data	Tamil-English	Malayalam-English
Training	27036	12057
Development	6759	3015
Test	8449	3768

Table 2
Class Distribution

	Sarcastic	Non-Sarcastic
Tamil-English-Train	7170	19866
Tamil-English-Dev	1820	4939
Malayalam-English-Train	2259	9798
Malayalam-English-Dev	588	2427

Malayalam-English code mixed data. The data set distribution is given in Table 1.

Table 2 shows the class-wise distribution of training and development sets of both languages. It is evident from the table that the data set is imbalanced.

4. Methodology

A few shot learning framework is used in our approach which is an efficient and prompt-free framework fine-tuned on sentence transformers [23]. This framework is built on sentence transformers, which are modified versions of pre-trained transformer models that create semantically significant phrase embeddings using Siamese and triplet network architectures. These models aim to increase the distance between sentence pairs that are semantically different and decrease the distance between pairs of sentences that are semantically similar. We have incorporated a 20-shot learning, in which Sentence Transformer is fine-tuned on 20 positive samples (Sarcastic) and 20 negative samples (Non-sarcastic) in a contrastive manner on sentence pairs. Both positive triplets i.e pairs of sentences randomly chosen from the same class and negative triplets i.e pairs of sentences randomly chosen from the different classes were used to fine tune the sentence transformers. These positive and negative triplets of both class labels namely Sarcastic and Non-sarcastic are concatenated and used for fine tuning the sentence transformers. After, fine tuning, the original training data was vectorized using sentence embeddings. These embeddings were used further to train a text classification head to determine whether a text is sarcastic or not. The process is shown in Figure 1.

SetFitTrainer¹ is used to implement our approach. Paraphrase-MiniLM-L3-v2 [24] embedding was used with logistic regression as a classification head to train our model.

¹https://huggingface.co/docs/transformers/main_classes/trainer#trainer

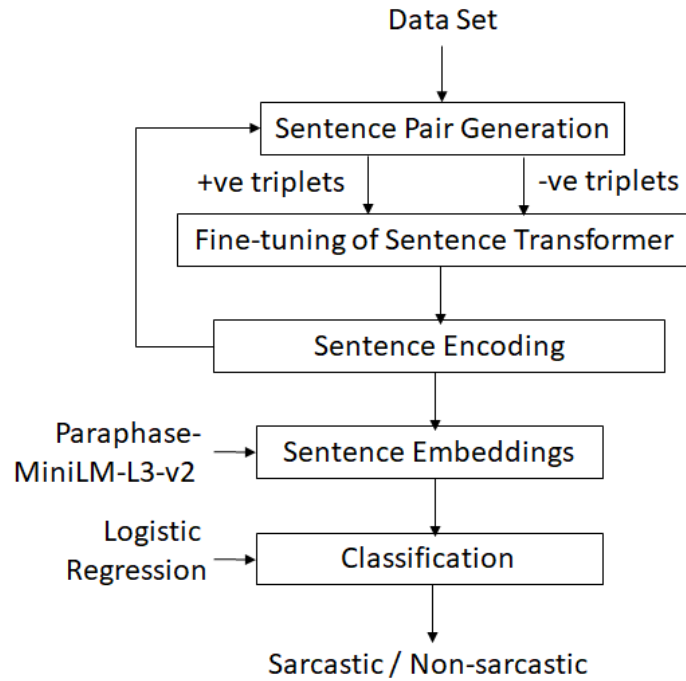


Figure 1: Methodology

5. Results and Performance Analysis

We have evaluated our few shot learning approach on Sarcasm_Identification_Draavidian-CodeMix@FIRE-2023 data set. We have also experimented three more models namely BERT, multilingual BERT and XLM-RoBERTa along with 2-way-20-shot learning. We have used metrics namely precision, recall and F1-score to evaluate the performance of our approaches.

Table 3 shows the performance of Tamil-English and Malayalam-English test data sets. Though BERT model performs better for Tami-English data set it is getting over-fit for Malayalam-English data set and gave a very low result for the recall.

6. Error Analysis

This section analyses some of the misclassifications in both Tamil-English and Malayam-English data sets. Our model wrongly predicts the Tamil-English sentence: “I support Dhraubathy, Nam naattil jaadhi madha veriyinar thirundhavendum, Nalla muyarchi, Brave attempt” as “Sarcastic”. The model learns the words like ”jaadhi” and ”madha” contributes to sarcasm from the training instances. However, the sentence is not sarcastic. Similary, a Malayalam-English sentence: “Mammootty nalla cinemakal vittu pinnem mass floppukalilekku pokanallo” which is sarcastic that has been classified as “Non-sarcastic” due to the terms “nalla” and “mass”. This is shown in Table 4.

Table 3
Performance Comparison on Test Set

Language	Model	Precision	Recall	F1-Score
Tamil-English	BERT	0.69	0.68	0.69
	XLNet	0.54	0.72	0.62
	mBERT	0.66	0.64	0.65
	2-way-20-shot	0.67	0.69	0.68
Malayalam-English	BERT	0.42	0.02	0.04
	XLNet	0.45	0.58	0.51
	mBERT	0.48	0.67	0.48
	2-way-20-shot	0.49	0.68	0.57

Table 4
Error Analysis

Instances	Predicted Labels	Original Labels
I support Dhraubathy, Nam naattil jaadhi madha veriyinar thirundhavendum, Nalla muyarchi, Brave attempt	Sarcastic	Non-sarcastic
Mammootty nalla cinemakal vittu pinnem mass floppukalilekku pokanallo	Non-sarcastic	Sarcastic

7. Conclusions

Identifying sarcasm is an important task in many applications such as sentiment analysis, hate speech detection, hate speech detection and homophobia identification. It is more challenging when the text is a code mixed. Several research works have been reported in English, Arabic and European languages for detecting sarcasm. However, it is in early stage in Dravidian languages. Sarcasm_Identification_Dravidian- CodeMix@FIRE-2023 shared task aims to address this problem by providing a data set to detect sarcasm in Tamil and Malayalam code mixed languages. We have implemented 4 models namely, BERT, mBERT, XLNet and 2-way-20-shot learning to detect sarcasm. 2-way-20-shot approach performs better for Malayalam-English data and for Tamil-English data it performs equal to BERT. Paraphrase-MiniLM-L3-v2 embeddings with logistic regression was used to train the model. In future, k values can be used in few shot learning for the better fine-tuning. Also, language agnostics embeddings can be used with other classifiers to improve the performance.

References

- [1] B. R. Chakravarthi, A. Hande, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, How can we detect homophobia and transphobia? experiments in a multilingual code-mixed

- setting for social media governance, *International Journal of Information Management Data Insights* 2 (2022) 100119.
- [2] B. R. Chakravarthi, Hope speech detection in youtube comments, *Social Network Analysis and Mining* 12 (2022) 75.
- [3] C. Techentin, D. R. Cann, M. Lupton, D. Phung, Sarcasm detection in native english and english as a second language speakers., *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 75 (2021) 133.
- [4] P. Katyayan, N. Joshi, Sarcasm detection approaches for english language, *Smart Techniques for a Smarter Planet: Towards Smarter Algorithms* (2019) 167–183.
- [5] K. A., T. D., Sarcasm identification and detection in conversation context using BERT, in: *Proceedings of the Second Workshop on Figurative Language Processing*, Association for Computational Linguistics, Online, 2020, pp. 72–76. URL: <https://aclanthology.org/2020.figlang-1.10>. doi:10.18653/v1/2020.figlang-1.10.
- [6] R. Justo, J. M. Alcaide, M. I. Torres, M. Walker, Detection of sarcasm and nastiness: new resources for spanish language, *Cognitive Computation* 10 (2018) 1135–1151.
- [7] S. K. Bharti, K. S. Babu, R. Raman, Context-based sarcasm detection in hindi tweets, in: *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)*, IEEE, 2017, pp. 1–6.
- [8] N. Pawar, S. Bhingarkar, Machine learning based sarcasm detection on twitter data, in: *2020 5th international conference on communication and electronics systems (ICCES)*, IEEE, 2020, pp. 957–961.
- [9] S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, M. Shrivastava, A corpus of english-hindi code-mixed tweets for sarcasm detection, *arXiv preprint arXiv:1805.11869* (2018).
- [10] K. Khandagale, H. Gandhi, Sarcasm detection in hindi-english code-mixed tweets using machine learning algorithms, in: *International Conference on Computing in Engineering & Technology*, Springer, 2022, pp. 221–229.
- [11] B. R. Chakravarthi, N. Sripriya, B. Bharathi, K. Nandhini, S. Chinnaudayar Navaneethakrishnan, T. Durairaj, R. Ponnusamy, P. K. Kumaresan, K. K. Ponnusamy, C. Rajkumar, Overview of the shared task on sarcasm identification of Dravidian languages (Malayalam and Tamil) in *DravidianCodeMix*, in: *Forum of Information Retrieval and Evaluation FIRE - 2023*, 2023.
- [12] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, K. Fu, Research progress on few-shot learning for remote sensing image interpretation, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021) 2387–2402.
- [13] D. Das, C. G. Lee, A two-stage approach to few-shot learning for image recognition, *IEEE Transactions on Image Processing* 29 (2019) 3336–3350.
- [14] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, J. Luo, Revisiting local descriptor based image-to-class measure for few-shot learning, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7260–7268.
- [15] D. Argüeso, A. Picon, U. Irusta, A. Medela, M. G. San-Emeterio, A. Bereciartua, A. Alvarez-Gila, Few-shot learning approach for plant disease classification using images taken in the field, *Computers and Electronics in Agriculture* 175 (2020) 105542.
- [16] L. Yan, Y. Zheng, J. Cao, Few-shot learning for short text classification, *Multimedia Tools and Applications* 77 (2018) 29799–29810.

- [17] R. Pasunuru, V. Stoyanov, M. Bansal, Continual few-shot learning for text classification, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5688–5702. URL: <https://aclanthology.org/2021.emnlp-main.460>. doi:10.18653/v1/2021.emnlp-main.460.
- [18] M. Hofer, A. Kormilitzin, P. Goldberg, A. Nevado-Holgado, Few-shot learning for named entity recognition in medical text, arXiv preprint arXiv:1811.05468 (2018).
- [19] C. I. Eke, A. A. Norman, L. Shuib, Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model, IEEE Access 9 (2021) 48501–48518.
- [20] A. Onan, Topic-enriched word embeddings for sarcasm identification, in: Software Engineering Methods in Intelligent Algorithms: Proceedings of 8th Computer Science On-line Conference 2019, Vol. 1 8, Springer, 2019, pp. 293–304.
- [21] S. Parveen, S. Saradha, N. Krishnaraj, An efficient detection and classification of sarcastic by using CNN model, in: Information Systems for Intelligent Systems: Proceedings of ISBM 2022, Springer, 2023, pp. 189–200.
- [22] R. Pandey, J. P. Singh, BERT-LSTM model for sarcasm detection in code-mixed social media post, Journal of Intelligent Information Systems 60 (2023) 235–254.
- [23] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, 2022. arXiv:2209.11055.
- [24] 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>.