Sarcasm Detection in Tamil and Malayalam Dravidian Code-Mixed Text

Supriya Chanda[1], Anshika Mishra[2] and Sukomal Pal[1]

[1]*Indian Institute of Technology (BHU), Varanasi, INDIA*
[2]*Vellore Institute of Technology Bhopal, Madhya Pradesh, INDIA*

Abstract

Sarcasm is a form of verbal irony that involves saying the opposite of what is actually meant in a mocking or humorous manner. You can find many sarcastic comments on social media these days, which are often code-mixed in nature.To gain insights from the textual data available to us, we need a system to detect sarcasm and identify the sentiments behind the texts. In this paper, we present a solution submitted for the shared task titled 'Sarcasm Identification of Dravidian Languages Tamil and Malayalam,' which was organized by Dravidian CodeMix 2023 at the Forum for Information Retrieval Evaluation (FIRE) 2023. This paper explores an approach to sarcasm detection, leveraging the BERT (Bidirectional Encoder Representations from Transformers) and a supplementary layer of neural networks for precise classification into two distinct classes: sarcastic and non-sarcastic comments. Our experiment demonstrates that our model effectively detects sarcastic comments, achieving an F1 score of 0.72 for both the Tamil-English and Malayalam-English code-mixed datasets.

Keywords

Social Media, Code-Mixed, BERT, Sarcasm, Sentiment Analysis, Tamil, Malayalam

1. Introduction

In our modern digital age, the complexities of human language present intriguing challenges for natural language processing systems. Among these intricacies, sarcasm emerges as a captivating linguistic puzzle. Sarcasm involves expressing thoughts in a manner that conceals the true intentions of the speaker, often infused with a dose of mockery or humor, serving as a linguistic tool to convey negative sentiments in a subtle manner. While humans excel at deciphering sarcasm through tone, context, and emotional cues, teaching machines to perform this feat remains a formidable undertaking.

Accurate detection of sarcasm holds significant importance, particularly in the domain of sentiment analysis, where it plays a pivotal role in understanding textual data. In our technology-dominated world, the proliferation of social media users has been nothing short of exponential, with a staggering 60% of the global population actively participating on these platforms, dedicating an average of 2 hours and 24 minutes daily to their online engagements

FIRE'23: Forum for Information Retrieval Evaluation, December 15-18, 2023, India

supriyachanda.rs.cse18@itbhu.ac.in (S. Chanda); anshika.mishra2019@vitbhopal.ac.in (A. Mishra); spal.cse@itbhu.ac.in (S. Pal)
https://cse-iitbhu.github.io/irlab/supriya.html (S. Chanda); https://cse-iitbhu.github.io/irlab/spal.html (S. Pal)
0000-0002-0877-7063 (S. Chanda); 0000-0001-8743-9830 (S. Pal)

(as reported by smartinsights [1]). Social media platforms provide an open canvas for individuals to openly express their views across a spectrum of subjects, events, personalities, and products, culminating in the generation of an astounding volume of data, estimated at a staggering 328.77 million terabytes daily. A substantial portion of this textual corpus is characterized by code-mixing, a linguistic phenomenon wherein individuals seamlessly blend elements from different languages, often employing the Roman script as a common bridge.

Code-mixing on social media mirrors the diverse linguistic backgrounds of users and the global reach of these digital platforms. It highlights how people consciously opt for code-mixing, skillfully weaving together different languages to enhance their communication, deftly switching between tongues to convey their feelings. Within the sphere of online interactions, encompassing comments, posts, and messages on social media platforms, the amalgamation of multiple languages, frequently expressed in the Roman script, is a common occurrence. The analysis of text not originally scripted in its native form presents an additional layer of complexity in the realm of natural language processing.

The importance of applying sentiment analysis to this data cannot be emphasized enough. It unveils a wealth of valuable insights, deeply influencing various fields such as market research, keeping a watchful eye on social media trends, and analyzing customer feedback. Additionally, it assumes a crucial role in countering the spread of hate speech on social platforms, thereby safeguarding the mental well-being of individuals. Furthermore, the capability to decode user queries infused with sarcasm paves the way for providing users with relevant information and responses.

The focal point of this shared task revolves around the precise identification of sarcasm and the determination of sentiment polarity within a code-mixed dataset comprising comments and posts in Tamil-English and Malayalam-English, all meticulously curated from the social media. It helps to explore how sarcasm is used in mixed-language conversations on social media. It's not just about language and computer challenges; it's also about gaining a deeper insight into how sarcasm works in the constantly changing world of online interactions. This, in turn, helps us become better at deciphering the subtleties of digital communication.

In this paper, we applied a method that leverages mBERT to enhance its capability in identifying sarcasm and determining sentiment polarity within code-mixed comments and posts written in Tamil-English and Malayalam-English, which are commonly encountered on social media platforms..

The rest of the paper is structured as follows. Section 2 provides a concise overview of prior research in this field. In Section 3, we delve into the datasets we utilized for our investigation. Section 4 elaborates on our computational methodologies, model specifications, and the techniques we employed for evaluation. Following this, we present our results and conduct a comprehensive analysis in Section 5. We conclude in Section 6.

[1] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

2. Related Work

Code-mixing in spoken language has been the subject of extensive research for decades. However, the analysis of code-mixed text, particularly in the context of social media, represents a relatively new frontier in the field of Natural Language Processing (NLP). Modern NLP models have demonstrated their prowess in various tasks, including sentiment analysis [1, 2, 3], language identification, hate speech identification [4, 5, 6], information retrieval [7], and named-entity recognition, particularly for monolingual text. Nevertheless, they face notable challenges when confronted with code-mixed content, where multiple languages are interwoven.

Sarcasm detection, a significant downstream task in the domain of NLP, has garnered considerable attention from researchers. Efforts have been directed toward effectively solving this challenging problem. One notable approach in this regard is the utilization of IndicBERT for detecting sarcasm in social media text, as proposed by Amir et al [8]. This model has captured contextual information and identifying sarcasm cues. Wicana et al. [9] undertook an extensive examination of sarcasm detection, delving into different machine-learning methods. Their review offers valuable perspectives on the latest techniques and the difficulties related to identifying sarcasm in text. They explored a range of neural network-based classification structures, including models like subword-level LSTM, Hierarchical LSTM, BERT, XLM RoBERTa, LSTM, GRU, and XLNet.

Notably, IndicBERT has proven to be proficient at capturing the nuances and language-specific characteristics of Indian languages [10]. To tackle the intricacies of sarcasm detection, researchers have employed a variety of techniques. For instance, an attention-based BiLSTM model, combined with a feature-rich Convolutional Neural Network (CNN) approach [11], has been utilized. It is essential to note that while sarcasm and hate speech are related, they are not the same, and they demand distinct approaches.

Efforts to identify sarcasm in social media content have sparked innovative approaches. One such method combines the multilingual BERT (mBERT) model with a Graph Convolution Network (GCN) [12]. Additionally, Agrawal et al. [13] ventured into the intriguing territory of utilizing emotional transitions to improve sarcasm detection, shedding light on the dynamic nature of emotional signals in recognizing sarcasm. Significantly, sarcasm detection goes beyond just written content. Pandey and Vishwakarma [14] tackled the challenge of multimodal sarcasm detection in videos. Their research focused on employing deep learning models to effectively leverage various sensory inputs, encompassing visual, audio, and textual cues, for sarcasm identification.

These studies collectively represent the evolving landscape of sarcasm detection, showcasing a wide array of approaches and techniques, with each offering valuable insights and advancements in the field.

3. Dataset

The dataset provided by the organizers is a valuable resource for our research, encompassing code-mixed comments and posts in Tamil-English and Malayalam-English, sourced from social media. While comments and posts may consist of multiple sentences, the dataset predominantly features an average sentence length of one. Importantly, each comment and post comes with sentiment polarity annotations, reflecting real-world scenarios and challenges associated with class imbalance. This dataset encourages us to investigate how sarcasm manifests in code-mixed contexts on social media. It includes development, training, and test dataset of YouTube video comments in Tamil-English and Malayalam-English, encompassing various code-mixing types and linguistic characteristics, providing a rich foundation for our research into sarcasm expression in these multilingual settings.

Table 1
Data Distribution for sarcasm detection of code-mixed text in Dravidian languages

Tamil - English			
Class	**Training**	**Development**	**Test**
Sarcastic	7170	1820	2263
Non-Sarcastic	19866	4939	6186
Malayalam - English			
Class	**Training**	**Development**	**Test**
Sarcastic	2259	588	685
Non-Sarcastic	9798	2427	3083

4. Methodology

4.1. Preprocessing

In the data preprocessing phase, we conducted several essential steps to refine the dataset. We removed hashtags, punctuation marks, URLs, numbers, and mentions that lacked clear semantic significance. Emojis were systematically replaced with their corresponding semantic text representations. Additionally, any extra white spaces or extra spaces were meticulously stripped from the dataset to ensure a clean and consistent text corpus for subsequent analysis.

4.2. Model Architecture

In our research methodology, we utilised the robust bert-base-multilingual-cased (mBERT) pre-trained models to create a solid foundation for our task. mBERT is built on the transformer architecture, which employs self-attention mechanisms both in the encoder and decoder. These models are pre-trained on vast text corpora, including Wikipedia, and have a well-established track record of delivering exceptional performance when fine-tuned for various downstream tasks.

For our specific objective of identifying code-mixed language and detecting sarcasm, we opted for the BERT (Bidirectional Encoder Representations from Transformers) model, with a focus on the multilingual variant known as mBERT (bert-base-multilingual-cased). This model's strength lies in its ability to handle text from a wide array of languages, featuring a substantial parameter count of 179 million, encompassing 12 transformer blocks, 768 hidden layers, and 12 attention heads.

Our architectural design begins with the model taking a special [CLS] token as input, followed by a sequence of words. This input traverses through the layers, with each layer applying self-attention mechanisms and forwarding the results to the subsequent encoder. The output from the final layer of the pre-trained mBERT model serves as the input to a softmax feedforward neural network, a critical component in classifying statements into two categories: Sarcastic or Non-Sarcastic. This neural network generates a probability distribution for each word within the sequence across predefined tags. During prediction, the tag with the highest associated probability is selected as the predicted tag for each word.

In the training phase, we carefully tuned specific hyperparameters to guide the learning process effectively. These included a learning rate of 0.01, a batch size of 16, and a maximum of 10 training epochs. These hyperparameters were meticulously optimized to ensure the model's proficiency in code-mixed language identification and sarcasm detection.

5. Results and Discussion

In this section, we delve into the comprehensive evaluation of our model's performance on both datasets: Tamil-English and Malayalam-English, as part of the Sarcasm Identification task within Dravidian Languages. The performance of our proposed models is examined using a range of evaluation metrics, with a primary focus on accuracy, recall, macro-averaged F1-score, and weighted average F1-score. The organizers thoughtfully provided test data for both Dravidian languages, which served as the foundation for our model evaluation.

Our methodology involved fine-tuning our model based on the training and validation datasets, ensuring it was well-prepared for the subsequent test data. Upon submission of our prediction file for the test data, we achieved an F1 Score of 0.72 for both language pairs. This score reflects a reasonable overall performance and places us at the third position in the ranking for both Tamil-English and Malayalam-English language pairs. Table 2 and 3 display the performance of the test outcomes for our proposed model and top scored team [15] for Tamil-English and Malayalam-English language respectively. Table 4 shows the class wise classification report for both language pairs on test data.

Table 2
F_1-scores for Tamil-English test data and rank list

Tamil - English		
Team Name	F_1 score	Rank
hatealert_Tamil	0.74	1 / 8
IRLabIITBHU_tam	0.72	3 / 8

While our system demonstrated commendable accuracy, it's worth noting that other competing teams surpassed us in both Precision and Recall, which ultimately influenced our F1 score and final ranking. This outcome encourages further refinement of our approach to enhance

Table 3

F_1-scores for Malayalam-English test data and rank list

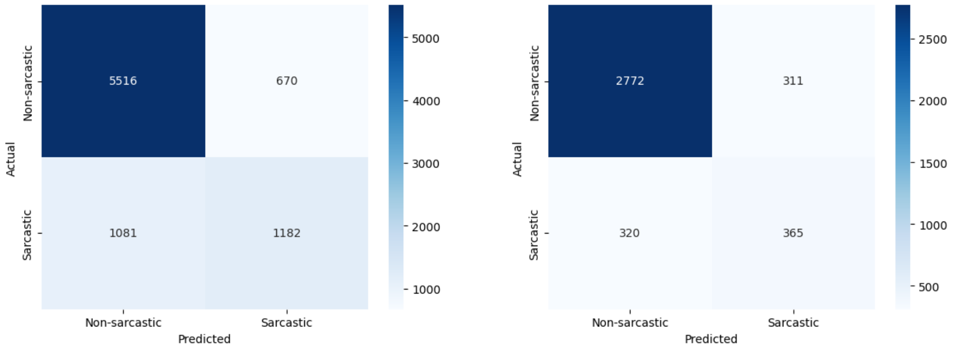
Malayalam - English		
Team Name	F_1 score	Rank
SSNCSE1_Malayalaml	0.74	1 / 8
IRLabIITBHU_mal	0.72	3 / 8

Table 4

Precision, recall, F_1-scores, and support for both language on test data

	Tamil - English				Malayalam - English			
	Precision	Recall	F_1-score	support	Precision	Recall	F_1-score	support
Non-sarcastic	0.84	0.89	0.86	6186	0.90	0.90	0.90	3083
Sarcastic	0.64	0.52	0.57	2263	0.54	0.53	0.54	685
macro avg	0.74	0.71	0.72	8449	0.72	0.72	0.72	3768
weighted avg	0.78	0.79	0.79	8449	0.83	0.83	0.83	3768
Accuracy	0.79			8449	0.83			3768

our model's precision and recall, aiming for even more competitive results in future endeavors. Figure 1 display the confusion matrices of two language pairs based on our subission.



(a) Tamil-English

(b) Malayalam-English

Figure 1: Confusion matrices for our submissions on the corpus test set for both language pairs (a) Tamil-English, (b) Malayalam-English

6. Conclusion

In this research, we've tackled the intricate task of identifying sarcasm and assessing sentiment polarity in code-mixed comments and posts, specifically in the Tamil-English and Malayalam-English languages, as extracted from the vibrant realm of social media. Our exploration into sentiment analysis reaffirms the growing significance of understanding user opinions, particularly in the context of enhancing business strategies. For our experimentation, we harnessed the capabilities of the pre-trained Multilingual BERT model, which yielded a commendable

F1 score of 0.72. This achievement reflects the effectiveness of our approach in capturing the nuances of sarcasm within code-mixed contexts. Despite our system's impressive accuracy, we acknowledge the competitive landscape where other teams excelled in both Precision and Recall, influencing our F1 score and final ranking. This outcome serves as a catalyst for refining our methodology further. In our future endeavors, we will be steadfast in our pursuit of enhancing precision and recall, with the aim of achieving even more competitive results.

References

[1] S. Chanda, S. Pal, Irlab@ iitbhu@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text., in: FIRE (Working Notes), 2020, pp. 535–540.

[2] S. Chanda, R. Singh, S. Pal, Is meta embedding better than pre-trained word embedding to perform sentiment analysis for dravidian languages in code-mixed text?, Working Notes of FIRE (2021).

[3] S. Chanda, A. Mishra, S. Pal, Sentiment analysis and homophobia detection of code-mixed dravidian languages leveraging pre-trained model and word-level language tag, in: Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR, 2022.

[4] A. Saroj, S. Chanda, S. Pal, Irlab@ iitv at semeval-2020 task 12: multilingual offensive language identification in social media using svm, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 2012–2016.

[5] S. Chanda, S. Ujjwal, S. Das, S. Pal, Fine-tuning pre-trained transformer based model for hate speech and offensive content identification in english, indo-aryan and code-mixed (english-hindi) languages, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.

[6] S. Chanda, S. Sheth, S. Pal, Coarse and fine-grained conversational hate speech and offensive content identification in code-mixed languages using fine-tuned multilingual embedding, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022.

[7] S. Chanda, S. Pal, The effect of stopword removal on information retrieval for code-mixed data obtained via social media, SN Computer Science 4 (2023) 494.

[8] S. Amir, B. C. Wallace, H. Lyu, P. Carvalho, M. J. Silva, Modelling context with user embeddings for sarcasm detection in social media, in: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 167–177. URL: https://aclanthology.org/K16-1017. doi:10.18653/v1/K16-1017.

[9] S. G. Wicana, T. Y. Ibisoglu, U. Yavanoglu, A review on sarcasm detection from machine-learning perspective, 2017 IEEE 11th International Conference on Semantic Computing (ICSC) (2017) 469–476. URL: https://api.semanticscholar.org/CorpusID:16074739.

[10] K. Jain, A. Deshpande, K. Shridhar, F. Laumann, A. Dash, Indic-transformers: An analysis of transformer language models for indian languages, 2020. arXiv:2011.02323.

[11] D. K. Jain, A. Kumar, G. Garg, Sarcasm detection in mash-up language using soft-attention

based bi-directional lstm and feature-rich cnn, Appl. Soft Comput. 91 (2020) 106198. URL: https://api.semanticscholar.org/CorpusID:216439240.

[12] M. Marreddy, S. R. Oota, L. S. Vakada, V. C. Chinni, R. Mamidi, Multi-task text classification using graph convolutional networks for large-scale low resource language, 2022. arXiv:2205.01204.

[13] A. Agrawal, A. An, M. Papagelis, Leveraging transitions of emotions for sarcasm detection, Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020). URL: https://api.semanticscholar.org/CorpusID:220729631.

[14] A. Pandey, D. K. Vishwakarma, Multimodal sarcasm detection (msd) in videos using deep learning models, in: 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), 2023, pp. 811–814. doi:10.1109/APSIT58554.2023.10201731.

[15] 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.