Sarcasm Detection in Dravidian Languages

Aioshi Chowdhury[1], Soumyajit Paul[1], Souvik Kundu[1], Anurag Kumar Thakur[1], Anuska Sarkar[1], Anusmita Ray Chaudhuri[1], Anirban Ray[1] and Debasree Mitra[2]

[1]*Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, India*
[2]*Department of Computer Science and Engineering, Adamas University, Kolkata, India*

Abstract

The big challenge in NLP for local Indian languages, mainly Tamil and Malayalam, is to develop a strong sarcasm detection system. The linguistic characteristics of these Dravidian languages are quite different, and there are very few annotated datasets. This makes sarcasm detection challenging because it relies on hints that aren't always obvious and depends on a person's understanding of the situation or topic. Hence, a model capable of capturing those aspects is required. To address the above problem, a deep learning-based model is proposed where Convolution layers take advantage of automatic feature extraction, max-pooling for reducing dimensions, and bidirectional GRUs to capture long-range dependencies and contextual information in sentences. The model performs very well in non-sarcastic and sarcastic detection, scoring an F1 of 0.94 for non-sarcastic detection 0.71 for sarcastic detection in Malayalam, and 0.94 for non-sarcastic and 0.81 for sarcastic detection in Tamil. This indicates it was able to handle the complexities of sarcasm detection in these low-resource languages successfully. Our rank for sarcasm detection in the Tamil language was 10, followed by rank 12 in the Malayalam language for the same.

Keywords

NLP, bidirectional GRUs, deep learning, linguistic characteristics, Dravidian languages,

1. Introduction

Sarcasm detection is a key part of sentiment analysis that plays an important role in the proper understanding of user sentiments and intentions on various digital platforms. Sentimental expressions, may take a rather direct form or may alter the meaning of a sentence significantly by deliberately altering its apparent meaning. This complexity makes it difficult for NLPs to recognize sarcasm with accurate classifications. This problem gets further complicated as Tamil and Malayalam languages are low-resourced languages and the different grammatical structures, vocabulary, and cultural nuances of Dravidian languages make it yet more complex. The data used is low-sourced and code-mixed and code-switched. The availability of fewer annotated datasets for sarcasm as well as diversity in expression forms a challenge to develop highly specialized models for sarcasm. All these challenges should be met to take regional language NLP capabilities forward and ensure improvements in the effectiveness of applications related to sentiment analysis. This research will develop a strong sarcasm detection system, targeted toward Tamil and Malayalam code-mixed and code-switched languages, with some highly advanced deep learning techniques to well capture the nuances of sarcasm existing in these languages. The proposed model uses Convolutional Neural Networks and MaxPooling to discover nature-like patterns followed by Bidirectional GRUs and LSTM layers in its attempt to improve the overall accuracy of sarcasm detection. Contributions will include enhanced performance metrics of sarcasm detection not only in Tamil but also in Malayalam, along with a deeper understanding of how best-advanced techniques from NLP can be adapted to low-resource languages.

Forum for Information Retrieval Evaluation, December 12-15, 2024, India
aioshichowdhury2004@gmail.com (A. Chowdhury); soumyajitpaul2002@gmail.com (S. Paul); souvikkundu1284@gmail.com (S. Kundu); iamanurag0708@gmail.com (A. K. Thakur); anuskasarkar06@gmail.com (A. Sarkar); titirray05@gmail.com (A. R. Chaudhuri); anirbanmark1429@gmail.com (A. Ray); debasree.mitra2005@gmail.com (D. Mitra)

2. Literature Survey

Rajnish Pandey and Jyoti Prakash Singh [1], effectively described techniques by using a hybrid model of BERT-LSTM for sarcasm detection in code-mixed social media posts for both English and Hindi languages. Meanwhile, "Multi-modal sarcasm detection and humor classification in code-mixed conversations" authored by Bedi, Manjot (2021) [2], introduces a multi-modal approach to detect sarcasm by using images, videos, or voices.

Aggarwal, and Akshita [3], showed addressing the focus on detecting sarcasm in Hindi-English code-mixed data by using bilingual word plantation among them.

Rosid, and Mochamad Alfan [4], introduced a research model that combines both the convolutional and the bi-directional GRU layers with a multi-headed attention mechanism for the detection of sarcasm in Indonesian-English code-mixed text. While Chanda, Supriya et al. [5], showed a method for detecting sarcasm in Tamil and Malayalam Dravidian code-mixed text using a combination of both classical machine learning and deep learning techniques by approaching the different challenges of Dravidian languages for both grammar and spelling checks for the informal way of communication for both Tamil and Malayalam.

Bhaumik, Anik Basu, and Das, Mithun [6], used numerous transformer-based models in code-mixed Tamil and Malayalam text to detect sarcasm. The usage of transformers activates the model to give better performance to intricate sentence structures present in code-mixed languages. This paper showcases the new challenge of identifying sarcastic comments and posts in code-mixed Dravidian languages.

Maity, Krishanu et al. [7], introduced a multitasking framework to detect sarcasm, emotion, and sentiment in code-mixed memes simultaneously. The framework joined text and image data to enhance detection accuracy in complex social media posts. To solve this task, they created a novel multi-modal meme dataset called MultiBully and showcased a new architecture which is called CLIP-CentralNet, which is an attention-based multi-task framework for sentiment, emotion, and sarcasm-aided cyberbullying detection.

Tejasvi, Koti, et al. [8], focused on traditional NLP techniques in Hindi-English code-mixed tweets to detect sarcasm. It speaks about the difficulties of using multiple languages in one context and suggests using a feature-based method to better detect sarcasm.

Shah, Aditya, and Maurya, Chandresh Kumar [9], examined discrepancies in detecting sarcasm in code-mixed text. The authors explain that the discrepancy between literal and intended meaning is a key indicator of sarcasm and demonstrate that it is effective in a code-mixed setting. This model is effective in capturing incongruity through FastText sub-word embeddings to detect sarcasm in the text.

Ratnavel, Rajalakshmi et al. [10], suggested the use of transformer models in code-mixed Tamil data to detect sarcasm. This model is better at understanding mixed-language sentences than existing techniques. In this paper, they tested four models: BERT, mBERT, XLM-RoBERTa, and 2-way-20-shot learning to identify sarcasm. The 2-way-20-shot approach works better for Malayalam-English data, while Tamil-English data performs similarly to BERT.

Bansal, Srijan, et al. [11], showcased how code-switching patterns are very impactful on the performance of NLP applications such as humor, sarcasm, and hate speech detection. They explained how these patterns can significantly improve model performance by identifying and utilizing them.

N. Sripriya et al. [12], provided an in-depth exploration of the Dravidian Language, particularly Malayalam and Tamil. The shared Task focuses on developing a dataset of code-mixed texts in Malayalam, English, and Tamil. On the other hand, Chakravarthi, and Bharathi Raja [13], investigated the problem of identifying hope speech detection which utters positive as well as supportive sentiment.

Chakravarthi, Bharathi Raja et al. [14] showed the typology, i.e., identifying hate speeches of the category Homophobia and Transphobia in code-mix. This posed an increased number of offensive languages towards the LGBTQIA+ community, and fewer methods focusing on its detection inspired us to create approaches — solutions to enhance social life among these communities. In future work, we target to optimize the loss functions and construct a multilingual Homophobia and Transphobia detection system for numerous languages. In this paper, we provide a novel rationale for the detection of sarcasm speech.

Chakravarthi, Bharathi Raja, et al [15] introduced their work to extract humor in the Dravidian language, mainly Tamil and Malayalam text The paper exon preprocessing techniques incorporating pre-trained models, in particular, BERT and distilBERT and conventional method SVM, TF-IDF.

Chakravarthi, et al [16] highlighted the linguistic complexities of these mixed-language texts and discussed methodologies through their work in their shared task at FIRE 2024 conference to improve sarcasm detection for better natural language understanding in Dravidian languages.

3. Dataset Description

The dataset from the DravidianCodeMix@FIRE-2024 shared task has been utilized for sarcasm detection, consisting of Tamil-English and Malayalam-English posts and comments. This dataset is code-mixed, meaning it includes a blend of words or phrases from both Tamil and English or Malayalam and English, within a single sentence or text. For prior context, Vijay, Deepanshu, et al. [17], introduced a dataset that is specifically created in Hindi-English code-mixed social media posts to detect irony. In this paper, they presented an annotated corpus of Hindi-English code-mixed text which consists of tweet IDs and the corresponding annotations. They have also presented a supervised system that is used for detecting irony in code-mixed text.

Suhaimin, Mohd Suhairi Md, et al. [18], focuses on issues within the public safety section in the dataset. Constructing this type of dataset that includes English and Malay code-mixing elements.

Swami, Sahil, et al. [19] created a collection for both English and Hindi to develop code-mixed tweets specifically for sarcasm detection. The paper highlights the different aspects of social media, which have numerous challenges in detecting sarcasm from a code-mixed language and the informal way of writing style of the public.

Chakravarthi, Bharathi Raja, et al. [20] showed an important benchmark dataset for sarcasm detection in code mixing in the Dravidian language.

Table 1 and 2 provided a summary of the key features of the datasets, presenting an in-depth look at the statistics for each of the three sets (Training, Development, and Test) in both languages. This breakdown offers a clear understanding of how the data is distributed and the size of the datasets, giving insight into the structure of the code-mixed texts used for this task.

Table 1
Stats of Tamil datasets

	TAM(Train)	TAM(Dev)	TAM(Test)
Total no. of sentences	29570	6336	6338
Total no. of words	322278	67595	69802
Max. length of sentences	113	152	95
Non-sarcastic to sarcastic ratio	2174:783	2315:853	4621:1717

Table 2
Stats of Malayalam datasets

	MAL(Train)	MAL(Dev)	MAL(Test)
Total no. of sentences	13188	2826	2826
Total no. of words	147675	31197	3079
Max. length of sentences	654	262	169
Non-sarcastic to sarcastic ratio	509:119	2305:521	1157:256

Our data was processed and annotated at the sentence level, with each sentence labeled as either sarcastic or non-sarcastic. To support model development and testing, the datasets for both Tamil and Malayalam were divided into three distinct sets: the training set, the development set, and the test set.

4. Methodology

This all-inclusive methodology allows for a strong approach in the building, training, and evaluation of a model capable of detecting sarcasm in both Malayalam and Tamil code-mixed and code-switched languages, using a combination of Convolutional and Recurrent Neural Networks and advanced preprocessing and evaluation techniques.

Text pre-processing is critical before feeding the text data to the model in order to make sure that the model can learn from the inputs at hand. The text is tokenized, and it thus transforms into sequences of integers, with each integer indicating one word from Malayalam or Tamil languages. Such sequences have been padded to equal uniform input sizes. Such uniformity is required by the model for consistent processing of data. This preprocess will make the model strong at handling variable lengths of sentences, as all input data would be in the expected format, thus helping detect sarcasm and allow generalization using multilingual text data.

The design is tailor-made for the specific nature of multilingual text data, including an embedding layer that transforms integer sequences into dense vectors and a Conv1D that captures local patterns in the data. For capturing both sides of dependencies that ultimately help in context understanding and nuance detection in complicated sarcasm, there is a bidirectional GRU layer. These representations are further refined by additional LSTM layers, and Dense layers having ReLU activation and Dropout have been added to prevent overfitting; the final Dense layer uses Sigmoid activation for binary classification. From Figure 1, it is observed that the process starts by loading the Malayalam and Tamil text data

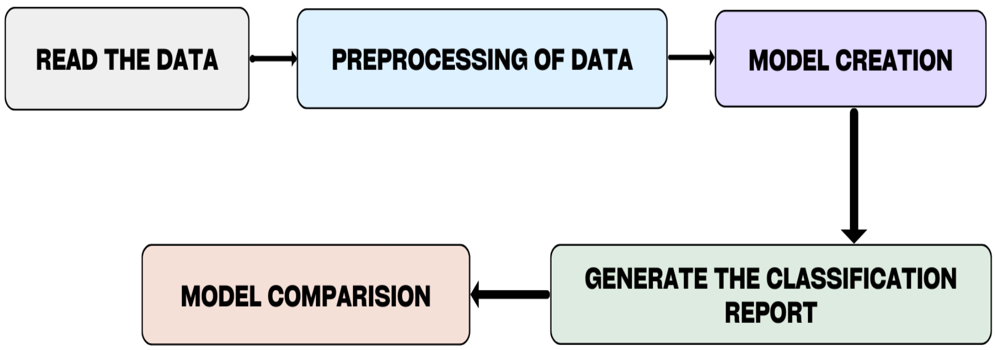

Figure 1: Procedure of our task.

mixed with both languages. Next, the text is broken down into numbers (tokenized) and padded to make sure all input has the same length. The numbers are turned into dense vectors in building the model, followed by layers that detect patterns and connections. Extra layers are added to prevent overfitting, and the final layer helps classify the text into two categories. After this, the models are compared, and a report is created to measure how well they perform using metrics like precision and F1 scores.

5. Result and Discussion

An analysis of the model shows that quite a lot is known about its performance for both Tamil and Malayalam. From the confusion matrix, it can be seen that the Malayalam model is successful in correctly classifying 2236 non-sarcastic instances and 323 sarcastic ones but fails to detect 69 non-sarcastic instances, misclassifying them as sarcastic, and 198 sarcastic instances as non-sarcastic, as shown in Figure 2.

On the other hand, the Tamil model correctly classified 4528 non-sarcastic and 1225 sarcastic instances, while labeling 102 non-sarcastic examples as sarcastic and missing 481 sarcastic examples, as indicated in Figure 3. These results are further depicted in the confusion matrices, where it can be visually observed how well the model performed in each category.

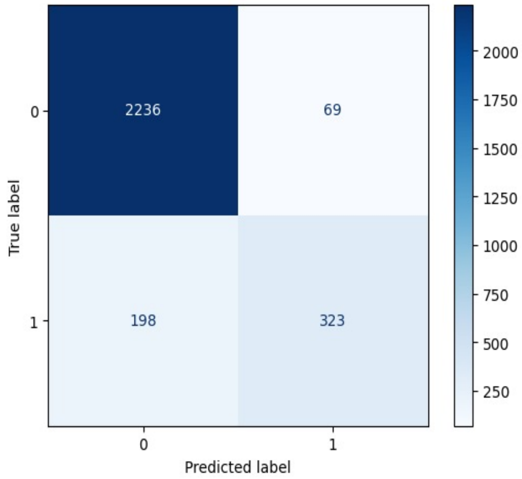

Figure 2: Confusion matrix for Malayalam language

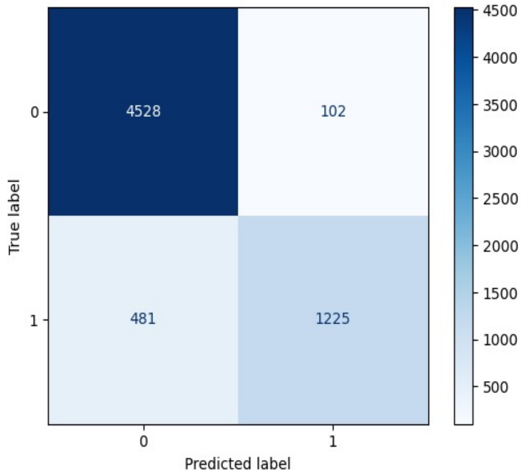

Figure 3: Confusion matrix for Tamil language

Tables 3 and 4 depict the F1 scores, where the Tamil model achieves the F1 score of 0.94 for non-sarcastic and 0.81 for sarcastic instances (Table 3) whereas the Malayalam model also achieves the F1 score of 0.94 for non-sarcastic and 0.71 for the sarcastic instances (Table 4). The graphical representation above gives a clear view of how well the model could identify sarcasm versus non-sarcasm across both languages.

Table 3
Classification Report for Tamil language

	Precision	Recall	F1-Score	Support
Non-sarcastic	0.90	0.98	0.94	4630
Sarcastic	0.92	0.72	0.81	1706
accuracy			0.91	6336
macro avg	0.91	0.85	0.87	6336
weighted avg	0.91	0.91	0.90	6336

Table 4
Classification Report for Malayalam language

	Precision	Recall	F1-Score	Support
Non-sarcastic	0.92	0.97	0.94	2305
Sarcastic	0.82	0.62	0.71	521
accuracy			0.91	2826
macro avg	0.87	0.80	0.83	2826
weighted avg	0.90	0.91	0.90	2826

Comparing the results, one can easily see that the results from both models are high in terms of accuracy for detecting non-sarcastic content, with an F1 of 0.94. It is quite evident that the Tamil model shows better detection of sarcastic material than the Malayalam model, with an F1 score of 0.81 as compared to 0.71. This shows that although both models perform similarly in correctly classifying non-sarcasm, the Tamil model tends to perform better in detecting sarcasm. The difference between the sarcasm detection systems calls for further development areas, especially the model of Malayalam because it will still need more improvement to increase its performance.

6. Conclusion and Future Scope

The sarcasm detection model shows a very good performance for both Tamil and Malayalam code-mixed texts, where both the languages showed a very good performance achieving a strong F1 score of 0.94 for detecting non-sarcastic content. However, there is a clear distinction in sarcasm detection, as the Tamil model outperforms the Malayalam model, with an F1 score of 0.81 compared to 0.71. While the model handles non-sarcastic classification efficiently, the gap in sarcastic content detection suggests room for improvement, particularly in the Malayalam model.

Looking ahead, future work could focus on developing a more robust fusion model that combines the strengths of Transformers, RNNs, and GRUs. Such a hybrid approach may better capture the nuances of sarcasm in code-mixed languages and improve overall detection accuracy. Incorporating these advanced architectures could address the current limitations, enhancing performance for both Tamil and Malayalam sarcasm detection tasks.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] 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.

[2] M. Bedi, S. Kumar, M. S. Akhtar, T. Chakraborty, Multi-modal sarcasm detection and humor classification in code-mixed conversations, IEEE Transactions on Affective Computing 14 (2021) 1363–1375.

[3] A. Aggarwal, A. Wadhawan, A. Chaudhary, K. Maurya, " did you really mean what you said?": Sarcasm detection in hindi-english code-mixed data using bilingual word embeddings, arXiv preprint arXiv:2010.00310 (2020).

[4] M. A. Rosid, D. Siahaan, A. Saikhu, Sarcasm detection in indonesian-english code-mixed text using multihead attention-based convolutional and bi-directional gru, IEEE Access (2024).

[5] S. Chanda, A. Mishra, S. Pal, Sarcasm detection in tamil and malayalam dravidian code-mixed text., in: FIRE (Working Notes), 2023, pp. 336–343.

[6] A. B. Bhaumik, M. Das, Sarcasm detection in dravidian code-mixed text using transformer-based models., in: FIRE (Working Notes), 2023, pp. 249–258.

[7] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1739–1749.

[8] K. Tejasvi, B. Reddy, S. Reddy, S. Rishikesh, Sarcasm detection for hindi english code mixed twitter data, International Journal for Research in Applied Science and Engineering Technology 11 (2023) 159–164.

[9] A. Shah, C. K. Maurya, How effective is incongruity? implications for code-mix sarcasm detection, arXiv preprint arXiv:2202.02702 (2022).

[10] R. Ratnavel, R. G. Joshua, S. Varsini, M. A. Kumar, Sarcasm detection in tamil code-mixed data using transformers, in: International Conference on Speech and Language Technologies for Low-resource Languages, Springer, 2023, pp. 430–442.

[11] S. Bansal, V. Garimella, A. Suhane, J. Patro, A. Mukherjee, Code-switching patterns can be an effective route to improve performance of downstream nlp applications: A case study of humour, sarcasm and hate speech detection, arXiv preprint arXiv:2005.02295 (2020).

[12] N. Sripriya, T. Durairaj, K. Nandhini, B. Bharathi, K. K. Ponnusamy, C. Rajkumar, P. K. Kumaresan, R. Ponnusamy, C. Subalalitha, B. R. Chakravarthi, Findings of shared task on sarcasm identification in code-mixed dravidian languages, FIRE 2023 16 (2023) 22.

[13] B. R. Chakravarthi, Hope speech detection in youtube comments, Social Network Analysis and Mining 12 (2022) 75.

[14] 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.

[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.

[16] B. R. Chakravarthi, S. N, B. B, N. K, T. Durairaj, R. Ponnusamy, P. K. Kumaresan, K. K. Ponnusamy, C. Rajkumar, Overview of sarcasm identification of dravidian languages in dravidiancodemix@fire-2024, in: Forum of Information Retrieval and Evaluation FIRE - 2024, DAIICT , Gandhinagar, 2024.

[17] D. Vijay, A. Bohra, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset for detecting irony in hindi-english code-mixed social media text., EMSASW@ ESWC 2111 (2018) 38–46.

[18] M. S. M. Suhaimin, M. H. A. Hijazi, E. G. Moung, Annotated dataset for sentiment analysis and sarcasm detection: Bilingual code-mixed english-malay social media data in the public security domain, Data in Brief 55 (2024) 110663.

[19] 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).

[20] B. R. Chakravarthi, N. Sripriya, B. Bharathi, K. Nandhini, S. C. 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.