Sarcasm Identification Of Dravidian Languages (Malayalam and Tamil)

V Indirakanth[1,**], Dharunkumar Udayakumar[2], Thenmozhi Durairaj[3] and B. Bharathi[4]

[1 2 3 4] *Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering*

Abstract

The rapid growth of social media has led to an increase in sarcastic comments. Detecting sarcasm in posts using multiple languages has become a critical aspect of language processing. Our work in the FIRE-2023 competition, specifically focused on 'Sarcasm Identification Of Dravidian Languages Tamil Malayalam', centered on identifying sarcasm in texts from social media, a research area of growing importance. We employed various models, including BERT, DistilBERT, XLM-RoBERTa, SVM, and TF-IDF, to categorize text as either sarcastic or not. Our team, SSN_FeaturesAlpha, achieved notable results, with the highest F1 score of 0.68 for Tamil using DistilBERT and 0.63 for Malayalam using BERT. It's worth highlighting that our submission for Tamil ranked 7th, and our Malayalam submission secured the 5th position, which underscores the effectiveness of our approach.

Keywords

Dravidian language, Text classification, Transfer learning, Sarcasm Detection

1. Introduction

In recent years, the volume of user-generated content on social media platforms like Twitter, Facebook, YouTube, and Instagram has grown exponentially. Social media is expected to continue being a major source of data in the years ahead. Sarcasm detection, in particular, has become a focal point in research due to its significance in moderating and comprehending online discourse. Detecting sarcasm involves the complex task of identifying sarcastic statements within text written in code-mixed form.

Tamil, an official Dravidian language spoken in Tamil Nadu (India), Sri Lanka, and Singapore, and Malayalam, another Dravidian language spoken in Kerala (India), are the languages under scrutiny. Code-mixing, where native speakers switch between multiple languages and sometimes employ the Roman script, is a common phenomenon in online social media interactions . The analysis of code-mixed bilingual or multilingual posts is gaining prominence in contemporary research. Identifying sentiments within indirect expressions such as sarcasm and metaphors is a challenging endeavor when done manually. Thus, the automatic detection of sentiments in various multilingual languages is a substantial challenge.

The paper with the title "Overview of The Shared Task on Sarcasm Identification of Dravidian Languages (Malayalam and Tamil) in Dravidian CodeMix"[9] provides an abstract view of the Shared Task on Sarcasm Identification in Dravidian Languages. The FIRE2023 competition, Sarcasm Identification of Dravidian Languages (Tamil and Malayalam), is designed to develop systems capable of discerning the polarity of sentiments in code-mixed Tamil and Malayalam languages within social media forums. The organizers of FIRE2022 initiated the task of identifying sentiments in Tamil and Malayalam code-mixed languages. Additionally, the challenge requires the categorization of posts into positive, negative, neutral, mixed emotions, or unclassifiable emotions in the intended language. This year's FIRE2023 competition provides datasets for two languages: Tamil and Malayalam, both mixed with English.

In the context of the FIRE2023 competition, which revolves around the challenge of detecting sarcasm in Tamil and Malayalam comments mixed with English, this article outlines our systematic approach. Our primary task was to categorize comments into positive, negative, unknown sentiment, mixed feelings, or unclassifiable in the respective language, ultimately determining if a comment is sarcastic. To achieve this, we followed a structured approach. Firstly, we implemented language-specific transliteration and translation techniques to enhance our understanding of the code-mixed content. Secondly, we conducted data preprocessing using the NLTK library for both training and test data in Tamil and Malayalam. Finally, we selected and fine-tuned a range of machine learning models, including BERT, DistilBERT, SVM, TF-IDF, and XLM-RoBERTa, integrating them with both Tamil and Malayalam while also incorporating English. This article provides a comprehensive overview of our methodology and contributions to the FIRE2023 competition, where our primary focus lies in the domain of sarcasm detection.

The structure of this paper is as follows: Section 2 reviews related work on sarcasm detection, Section 3 provides a detailed description of the data and our model methodology, Section 4 presents our experimental results and analysis, and finally, Section 5 offers conclusions drawn from our work and discusses potential avenues for further improvement in sarcasm detection.

2. Related Work

In recent years, the field of Natural Language Processing (NLP) has witnessed a surge of interest in the detection of sarcasm, a form of figurative language that presents a formidable challenge due to its often subtle and context-dependent nature. While the majority of research in this domain has focused on widely spoken languages such as English, there is a growing recognition of the need to extend this investigation to underrepresented languages like Tamil and Malayalam, which belong to the Dravidian language family. Previous work on sarcasm detection in NLP has predominantly leveraged various machine learning techniques, including supervised, unsupervised, and deep learning approaches. In the paper [3], they created a multilingual dataset to recognize and encourage positivity in the comments and proposed a novel custom deep network architecture that uses a concatenation of embeddings from T5-Sentence. They have experimented with multiple machine learning models, including SVM, logistic regression, K-nearest neighbour, and decision trees. The paper [4] aims at developing a system that groups posts based on emotions and sentiment and finds sarcastic posts, if present. The proposed system is to develop a prototype that helps come to an inference about the emotions of the

posts, namely anger, surprise, happiness, fear, sorrow, trust, anticipation, and disgust, with three sentic levels in each. The task [5] presents the findings of the shared task on Multimodal Sentiment Analysis and Troll meme classification in Dravidian languages. This task assumes the analysis of both textual and image features for making better predictions. The paper [6] investigates negative sentiment tweets with the presence of hyperbole for sarcasm detection. Six thousand and six hundred pre-processed negative sentiment tweets comprising Chinesevirus, Kungflu, COVID19, Hantavirus and Coronavirus were gathered for sarcasm detection. In paper [7] tasks included code-mixing at the intra-token and inter-token levels. In addition to Tamil, Malayalam and Kannada were also introduced. the quality and quantity of the submission show that there is great interest in Dravidian languages in code-mixed setting and state of the art in this domain still needs improvement. Task [8] intends to improve offensive language identification by generating pseudo-labels on the dataset. A custom dataset is constructed by transliterating all the code-mixed texts into the respective Dravidian language, either Kannada, Malayalam, or Tamil and then generating pseudo-labels for the transliterated dataset. The two datasets are combined using the generated pseudo-labels to create a custom dataset called CMTRA. As Dravidian languages are under-resourced, their approach increases the amount of training data for the language models.

## 3. Dataset Description and Proposed Methodology

This section provides information about the mixed-language data in Tamil and Malayalam, including details about the dataset and how we prepared it. In our research, we explored various techniques commonly used in Natural Language Processing (NLP). We used methods like Support Vector Machines (SVM), DistilBERT, XLM-RoBERTa, and transfer learning to improve our results.

3.1. Data Description

For the task of offensive detection, the organizers offered datasets that were code-mixed in Tamil and Malayalam. The Malayalam dataset has 12,057 posts for training and 3,768 posts for testing, whereas the Tamil dataset has 27,036 posts for training and 8,449 posts for testing the model system. The objective of this work is to divide the postings into two categories in the datasets for Tamil and Malayalam: Sarcastic and Non-sarcastic (Fig 1 & 2). The training data was taken from comments on YouTube [1] [2].

3.2. Data preprocessing

In the context of our research paper, data preprocessing was conducted on both the Tamil and Malayalam datasets to ensure their adaptability for sarcasm detection tasks in Dravidian languages. This preprocessing, executed using NLTK2, aimed to enhance the quality and uniformity of the text data. Initially, duplicate entries were removed to mitigate their potential influence on model performance. Subsequently, text strings beginning with "@" symbols, typically representing author names or user IDs, were eliminated. Hashtags, punctuation, URLs, and numerals devoid of semantic significance were also stripped from the text. Emojis were removed
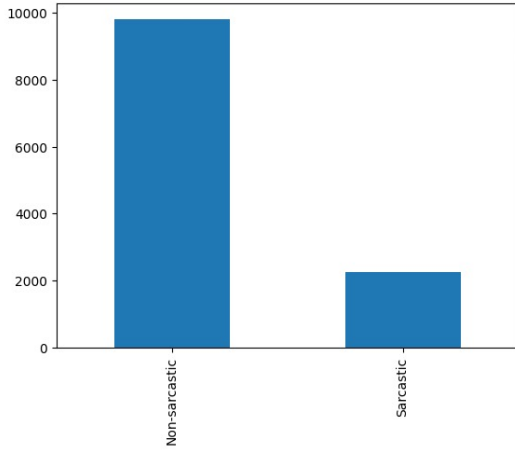
Figure 1: Tamil Training data values
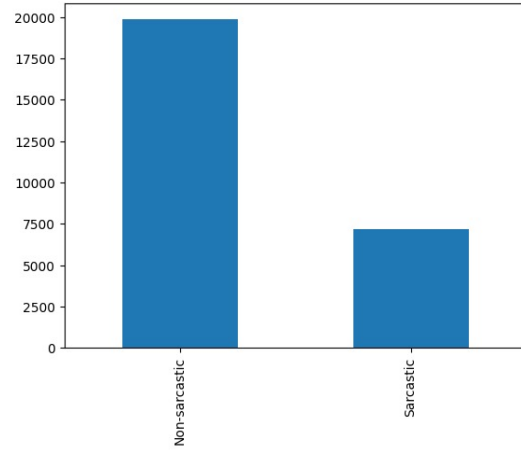
Figure 2: Malayalam Training data values

to maintain textual clarity. Additionally, all uppercase text in English and native language text in the Roman script were converted to lowercase. These preprocessing measures were consistently applied to both the Tamil and Malayalam datasets, ensuring the data's readiness for sarcasm detection tasks within the Dravidian language context. We downsampled training data, removed emojis, and eliminated common stop words to enhance data quality. We also meticulously fine-tuned labels for clarity, transforming "Sarcastic" and "Non-sarcastic" into "1" (sarcasm) and "0" (non-sarcasm), ensuring that our data aligns perfectly with our binary classification task.

Table 1
Preprocessing Results

	Text
Before Preprocessing	��������� enaku matum Tha apdi thonutho... ??? ADI THOOL
After Preprocessing	enaku matum tha apdi thonutho adi thol

3.3. Proposed Methodology

In this section, we delve into the methodology employed for the challenging task of sarcasm detection in Dravidian languages. Our aim is to break down the intricate process, highlighting its various stages, and elucidate how each step contributes to our overarching goal. Sarcasm, known for its subtlety and context-dependency, presents a formidable challenge in the realm of Natural Language Processing (NLP). Our approach leverages NLP techniques and a range of machine learning models to tackle this linguistic puzzle.

Tokenization:
For the initial data preprocessing, we utilized tokenization to break down the text into individual tokens or subwords. Depending on the model, different tokenizers were employed, such as the BertTokenizer for BERT and DistilBERT, XLMRobertaTokenizer for XLM-RoBERTa, and

TfidfVectorizer for the Support Vector Machine (SVM) approach. Tokenization is a crucial step in preparing the text data for further processing.

Model Training:

Next, we dive into the heart of our methodology: model training. Our ensemble of models, comprising BERT, DistilBERT, and XLM-RoBERTa, embarks on an extensive training journey. These models come pre-equipped with vast textual knowledge, but to excel at the specific task of sarcasm detection in Dravidian languages, they undergo fine-tuning. The fine-tuning process serves as the bridge that adapts these models to the nuanced world of Dravidian sarcasm. Fine-tuning involves training for 4 epochs, with each epoch representing a complete pass through the training data. Reprocessing input data is enabled to adapt the models to sarcasm detection characteristics. Additionally, the models are configured for binary classification, distinguishing between sarcastic and non-sarcastic text. This comprehensive fine-tuning process ensures that our models effectively capture the subtleties of sarcasm, enhancing their performance on the specific sarcasm detection task in Dravidian languages.

SVM (Support Vector Machine):

Our methodology extends its reach beyond neural network-based models. We introduce the venerable Support Vector Machine (SVM), a classic machine learning algorithm, into the mix. SVM stands as a benchmark for comparison with our neural network counterparts. Leveraging TF-IDF vectors as features and a linear kernel for classification, SVM offers a different perspective on sarcasm detection, enriching our research toolkit.

Model Evaluation:

To gauge the effectiveness of our sarcasm detection models, we embark on a comprehensive evaluation journey. Our evaluation relies on standard metrics that provide insight into each model's performance.

Classification Report: Our classification report emerges as a comprehensive mirror, reflecting metrics like accuracy, F1-score, precision, and recall. It offers a detailed assessment of our models, shedding light on their prowess in distinguishing between the worlds of sarcasm and non-sarcasm.

Confusion Matrix: In addition to the classification report, we construct confusion matrices that bring our model's classification results to life. These matrices vividly portray true positives, true negatives, false positives, and false negatives, painting a clear picture of our models' strengths and areas for improvement.

Prediction and Labeling:

Our methodology culminates in the real-world application of our trained models. We unleash them on testing data, stored conveniently in a CSV file. To facilitate interpretation, we introduce a new column labeled "Labels" into the data. Within this column, values of 0 denote "Non-sarcastic," while values of 1 signify "Sarcastic" based on our model's astute predictions.

In essence, our methodology represents a harmonious blend of NLP techniques and a diverse range of models. Together, tokenization, model training, and meticulous evaluation allow us to unravel the intricacies of sarcasm detection in Dravidian languages. Through this journey, we seek to gain insights into the effectiveness of each approach and distill meaningful conclusions about the task at hand.

4. Results

In this section, we present the evaluation of our model and submitted results for the Tamil, and Malayalam code-mixed languages.

4.1. Experimental Results

The below given results are the F1 Scores for each models used. The highest score recorded for Tamil language is 0.75 with DistilBERT model and highest score recorded for Malayalam language is 0.72 with BERT model. The confusion matrix provided is the validation / development data.

Table 2
Classification Report - BERT Malayalam

Title	Precision	Recall	F1 score	Support
0	0.93	0.70	0.80	2427
1	0.38	0.77	0.51	588
accuracy			0.72	3015
macro avg	0.66	0.74	0.66	3015
weighted avg	0.82	0.72	0.74	3015

Table 3
Classfication Report - DistilBERT Tamil

Title	Precision	Recall	F1 score	Support
0	0.90	0.75	0.82	4939
1	0.53	0.77	0.63	1820
accuracy			0.75	6759
macro avg	0.71	0.76	0.72	6759
weighted avg	0.80	0.75	0.77	6759

Table 4
Results

Model	Tamil	Malayalam
BERT	0.70	0.72
DistilBERT	0.75	0.64
XLM-RoBERTa	0.64	0.63
SVM	0.74	0.71
TF-IDF	0.72	0.70

4.2. Submitted Results

We applied the transfer learning technique to enrich the training data for the two languages. Our approach utilized a combination of BERT, DistilBERT, XLM-RoBERTa, SVM, and TF-IDF models.
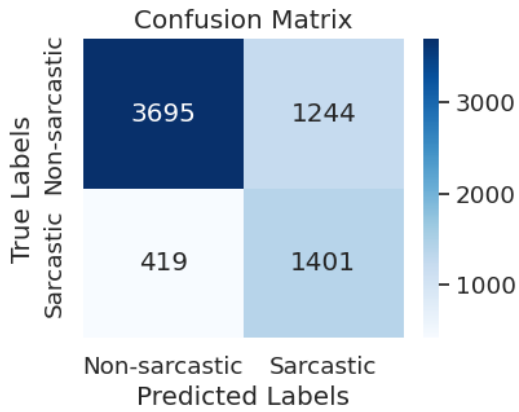
Figure 3: DistilBERT model for Tamil   **Figure 4:** BERT for Malayalam
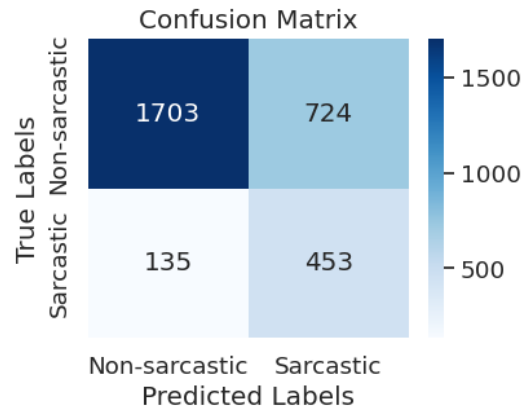
Development data Confusion matrix results

The collaborative efforts of our team, FeaturesAlpha_tam, yielded impressive results, with the maximum F1-scores reaching 0.68 in Tamil and 0.63 in Malayalam using various models.

In our submissions, we presented the results of DistilBERT for Tamil and BERT for Malayalam, achieving commendable rankings of 7th for Tamil and 5th for Malayalam in the competition. Table 5 represents Tamil Ranking and Table 6 represents Malayalam ranking. Detailed performance metrics, including accuracy and F1-scores, are available in Table 3 for Tamil and Table 2 for Malayalam, providing a comprehensive view of our models' effectiveness.

S.No	Team name	MF1	Rank
1	hatealert_Tamil	0.74	1
2	ABC_tamil	0.73	2
3	SSNCSE1_Tamill	0.73	2
4	IRLabIITBHU_tam	0.72	3
5	ramyasiva_tamil	0.71	4
6	ENDEAVOUR	0.70	5
7	MUCS_Tamil	0.70	5
8	Hydrangea_tamilrun3	0.69	6
9	**SSN_FeaturesAlpha_tam**	0.68	7
10	YenCS_tam	0.68	7
11	TechWhiz_tam	0.66	8

Table 5
Tamil Ranking

Furthermore, we've presented the prediction values in confusion matrices for both languages, enhancing the interpretability of our results. These visualizations can be found in Figure 3 for Tamil and Figure 4 for Malayalam, allowing for a deeper understanding of the model's performance.

S.No	Team name	MF1	Rank
1	SSNCSE1_Malayalaml	0.74	1
2	hatealert_Malayalam	0.73	2
3	ABC_malayalam	0.72	3
4	IRLabIITBHU_mal	0.72	3
5	MUCS_mal	0.71	4
6	**SSN_FeaturesAlpha_mal**	0.63	5
7	TechWhiz_mal	0.63	5
8	YenCS_mal	0.63	5
9	Hydrangea_malayalamrun1	0.57	6
10	ENDEAVOUR_malayalam	0.53	7
11	ramyasiva_malayalam	0.52	8

Table 6
Malayalam Ranking

5. Conclusion

This paper outlines our methodology for identifying sarcasm in text from Dravidian languages, specifically Tamil and Malayalam. Our approach focused on preprocessing techniques and the utilization of pre-trained models such as BERT, DistilBERT, XLM-RoBERTa, as well as traditional methods like SVM and TF-IDF, with various input variations for the shared task across both languages. Through rigorous evaluation, our findings indicate that fine-tuning the BERT and DistilBERT architecture yields notable performance improvements. Our team achieved enhanced F1 scores compared to baseline scores, showcasing the effectiveness of our approach. Additionally, we harnessed the power of transfer learning to maximize results. While our current research has shown promising results, there remains room for further advancements. Future research endeavors can explore the potential of different deep learning algorithms to push the boundaries of sarcasm detection in Dravidian languages. Moreover, extending our work to include other languages promises to broaden the scope and applicability of our methodology. In conclusion, our study represents a significant step forward in the realm of sarcasm detection within Dravidian languages. By comparing our methodology and outcomes with existing research, we hope to contribute to the ongoing dialogue and innovation in this field, ultimately paving the way for more accurate and robust sarcasm detection systems.

6. References

[1] Chakravarthi, B.R. 2022. Hope speech detection in YouTube comments. Social Network Analysis and Mining, 12(1), 75. Springer.

[2] Chakravarthi, B.R., Hande, A., Ponnusamy, R., Kumaresan, P.K., and Priyadharshini, R. 2022. 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(2), 100119. Elsevier.

[3] Chakravarthi, B.R. Hope speech detection in YouTube comments. *Soc. Netw. Anal. Min.* 12, 75 (2022). https://doi.org/10.1007/s13278-022-00901-z

[4] S. Rendalkar and C. Chandankhede, "Sarcasm Detection of Online Comments Using Emotion Detection," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2018, pp. 1244-1249, doi: 10.1109/ICIRCA.2018.8597368. URL:https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=8597368isnumber=8596764

[5] Premjith B, Chakravarthi BR, Subramanian M, et al. Findings of the Shared Task on Multimodal Sentiment Analysis and Troll Meme Classification in Dravidian Languages. Association for Computational Linguistics. January 2022. doi:10.18653/v1/2022.dravidianlangtech-1.39

[6] A Machine Learning Approach in Analyzing the Effect of Hyperboles Using Negative Sentiment Tweets for Sarcasm Detection." A Machine Learning Approach in Analyzing the Effect of Hyperboles Using Negative Sentiment Tweets for Sarcasm Detection - ScienceDirect, 22 Jan. 2022, doi:10.1016/j.jksuci.2022.01.008.

[7] Priyadharshini R, Chakravarthi BR, Thavareesan S, Chinnappa D, Thenmozhi D, Ponnusamy R. Overview of the DravidianCodeMIX 2021 Shared Task on sentiment Detection in Tamil, Malayalam, and Kannada. Forum for Information Retrieval Evaluation. December 2021. doi:10.1145/3503162.3503177

[8] Hande, Adeep, et al. "Offensive Language Identification in Low-resourced Code-mixed Dravidian Languages Using Pseudo-labeling." arXiv.org, 27 Aug. 2021, arxiv.org/abs/2108.12177v1.

[9] 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.