Transactions on Asian and Low-Resource Language Information Processing

An Empirical Study on Large Language Models for Sarcasm Detection in Low-resourced Indic Languages

Journal:	*Transactions on Asian and Low-Resource Language Information Processing*
Manuscript ID	TALLIP-24-0520
Manuscript Type:	Short Paper
Date Submitted by the Author:	06-Aug-2024
Complete List of Authors:	Desai, Dr. Mitali; Sarvajanik College of Engineering and Technology, Information Technology Gajjar, Khushi; CK Pithawalla College of Engineering and Technology Raval, Meetkumar; CK Pithawalla College of Engineering and Technology
Keywords:	Large Language Models, Multi-lingual Data, Sarcasm Detection, India Regional Languages, Natural Language Processing

An Empirical Study on Large Language Models for Sarcasm Detection in Low-resourced Indic Languages

MITALI DESAI*, Sarvajanik College of Engineering and Technology, India
KHUSHI GAJJAR† and MEET RAVAL†, C. K. Pithawala College of Engineering and Technology, India

Sarcasm is a commonly observed linguistic form for criticism. It often changes the polarity of the context present in the text; therefore, detecting sarcasm is considered as one of the significant challenges in various Natural Language Processing (NLP) tasks. The existing literature discloses the need to address the growing demand for the accurate and efficient detection of sarcasm in online content across multiple Indian languages. This study focuses on the exclusive linguistic characteristics of Indic low-resourced languages - Bengali, Hindi, Gujarati, Marathi, Punjabi, Telugu and Tamil to develop a comprehensive solution for multi-lingual sarcasm detection. By creating custom datasets of 60,000 samples of different Indian regional languages and leveraging the linguistic landscape of India, this study seeks to evaluate the performance of cutting-edge large language pre-trained models for the sentence-level sarcasm detection. The findings of this study have the potential to contribute in enhancing an overall capability of NLP systems in understanding and interpreting sarcasm across a diverse linguistic context of India.

CCS Concepts: • **Information systems** → *Sentiment analysis*; **Data extraction and integration**; **Structure and multilingual text search**.

Additional Key Words and Phrases: Large Language Models, Multi-lingual Data, Sarcasm Detection, India Regional Languages, Natural Language Processing

ACM Reference Format:
Mitali Desai, Khushi Gajjar, and Meet Raval. 2024. An Empirical Study on Large Language Models for Sarcasm Detection in Low-resourced Indic Languages. 1, 1 (August 2024), 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Social media platforms have become a convenient medium for users to share their opinions about any imminent events and/or feedback about the past events. Each user has a potential to extend his/her opinions (positive or negative) across the social media communal [14]. This user generated content is highly impactful in applications such as recommendation systems, sentiment analysis, opinion mining, influence identification and community analysis [13]. Detecting sarcasm in the user generated content on various social media platforms is one such compelling application [19].

Sarcasm is defined as a form of satire that is intended to express criticism or ridicule the opinions and/or ideas expressed by others [32]. Sarcasm can be expressed in textual, verbal and even in pictorial forms. When sarcasm is expressed in a textual form, it is difficult to be identified by a common person due to the absence of tone and/or gesture in text data [11].

Authors' Contact Information: Mitali Desai, mitalidesai17@gmail.com, Sarvajanik College of Engineering and Technology, Surat, Gujarat, India; Khushi Gajjar, khushigajjar456@gmail.com; Meet Raval, ravalmeett@gmail.com, C. K. Pithawala College of Engineering and Technology, Surat, Gujarat, India.

ACM XXXX-XXXX/2024/8-ART
https://doi.org/XXXXXXX.XXXXXXX

This study focuses on textual form and specifically, atomic sentences. In this case, sarcasm is often represented as a positive sentiment on surface level; however, the underlying context of the statement is negative. As an example, the sentence "Nice perfume. How long did you marinate in it?" seems to be a positive statement on the surface; however, is sarcastic in the actual context. Due to this figurative nature of sarcasm, it is considered as a crucial aspect in various Natural Language Processing (NLP) tasks [25]. Sarcasm detection is defined as a computational approach that automatically classifies whether the given text is sarcastic [41].

For automatic sarcasm detection in text data, various approaches including lexicon, rule-based NLP, pattern-based, corpus-based, statistical and machine learning are found in the literature [17]. In recent times, deep learning-based approaches, mainly, Large Language Models (LLMs) are widely employed in this domain [1, 8, 18, 24, 34, 37, 40].

Several studies have been published focusing on sarcasm detection in different languages. Majority of these studies have focused primarily on the English language whereas several studies on other languages such as Arabic, Chinese, Italian, Dutch, Greek and Indonesian have also been reported [19]. On the other hand, few recent studies have been conducted with the focus on Indian regional languages such as Hindi [7, 10, 20, 22, 39], Tamil [4, 23, 38], Telugu [9, 23, 33, 38], Marathi [31], Punjabi [3] and Bengali [27, 29]. It is observed that all these languages including Gujarati have recently gained the focus of researchers in the domain of sarcasm detection. The major reason is that many official languages of India are still low resource or no resource when it comes to lexical resources [6]. Hence, sarcasm detection becomes a tedious task.

Taking the above notion into consideration, this study attempts to empirically analyze various extremely utilized LLMs - mBERT, indicBERT, DistilBERT, MuRIL, mT5 and Bloom on low-resourced Indic Languages - Bengali, Hindi, Gujarati, Marathi, Punjabi, Telugu and Tamil languages. The mentioned LLMs are implemented on each language individually and on a consolidated multilingual dataset. To perform this analysis, we have collected data samples for every focused language from various online platforms and subsequently, curated a dataset consisting of 60,000 data samples. The collected data is pre-processed and then, the LLMs are implemented on i) an individual dataset of each focused language and ii) the multilingual dataset consisting of data samples for each focused language. The standard measures such as accuracy, precision and recall are considered to evaluate the performance. It is observed that mBERT, MuRIL and mT5 perform efficiently over other LLMs while considering language specific datasets. On the other hand, MuRIL outperforms other LLMs for multilingual datasets. The Outcome of this study provides a clear direction into further exploring the transformer based LLM models for diverse low-resources Indic languages. Also, the curated dataset of this study will be utilized for relevant studies in this domain.

The remainder of this paper is organized as follows: In Section 2, the existing literature is depicted with respect to sarcasm detection in text data. In Section 3, the methodology to conduct this study is explained in detail. In Section 4, the result analysis is presented. In Section 5, this study is concluded with the future expansion possibilities.

2 Literature Review

Sarcasm is form of verbal satire that is intended to express criticism or ridicule the opinions and/or ideas expressed by others. In visual forms of media such as videos, sarcasm can be determined through body language and facial expressions whereas for auditory media, changes in tone serve as cues for sarcasm detection. In both cases evaluating the context and tone allow to ascertain whether the expression is sarcastic or not. However, when expressed in the text format, detecting sarcasm efficiently becomes challenging [11].

A new dataset for sarcasm detection was introduced by Misra and Arora [28] utilizing the sarcastic news website and real news website. The authors highlighted the limitations such as

noisy labels or limited scale in the currently utilized datasets for sarcasm detection. On the other hand, the researchers provided an extensive dataset emphasizing high-quality labels to enhance sarcasm detection through deep learning models. They achieved promising outcomes, showcasing the effectiveness of their dataset and hybrid neural network architecture approach.

The authors explored the complex challenge of detecting sarcasm in text data, emphasizing on the difficulties arising from its reliance on context [35]. The research employs a dataset consisting of 1.3 million social media comments, covering both sarcastic and non-sarcastic content. The authors have employed machine learning models such as logistic regression, ridge regression and support vector machines alongside deep learning models like Bidirectional Long Short-Term Memory (BiLSTM) and a Bidirectional Encoder Representations from Transformers (BERT) model to address the challenges involving preprocessing of data and extraction of features through natural language processing methods. The findings of this research reveal that deep learning models, specifically the BERT-based model, demonstrate superior performance in identifying sarcasm in social media comments compared to conventional machine learning methods.

The authors in [21] have investigated the significance of context on sentiment analysis, focusing specifically on user-generated content from various social media platforms. Three predictive learning models i) a traditional Term Frequency-Inverse Document Frequency (TF-IDF) approach with ensemble voting, ii) a combination of semantic and pragmatic features with various classifiers and iii) a deep learning model using Bi-directional LSTM with GloVe word embeddings are implemented on Twitter and Reddit datasets. It was revealed through the empirical study that BiLSTM model dominated with the accuracy of 86.32% and 82.91% for Twitter and Reddit datasets, respectively.

A study on automatic sarcasm detection was presented in [19] to examine the importance of sentiment analysis. Through their research on various characteristics in the field of sentiment analysis, the authors demonstrated the importance of comprehending the nuanced forms of sarcasm as well as the challenges involved in its computational linguistics. Finally, the authors disclosed an effective roadmap for more research in this field.

The authors in [5] have emphasized on the pivotal role of contextual linguistic cues while detecting sarcasm in Twitter data. While comparing with the traditional methods that prioritize contextual information such as environmental factors and the relationship between speakers and their audiences, the authors described sarcasm as inherently dependent on context, requiring shared understanding between speakers and their audiences. The proposed method leads to improved accuracy in sarcasm detection by considering the interpersonal interaction within the conversations.

In [34], the authors performed sarcasm detection in Tweeter data by employing deep learning models with manually crafted contextual features. The features extracted from Convolutional Neural Network (CNN) are aggregated with context specific handcrafted features. The aim was to identify optimal features from the independent and aggregated feature sets. The outcome revealed that logistic regression method had effective results among others.

The authors emphasized on Pre-Trained Language (PLMs) models for multi-class text classification in financial domain [2]. The authors suggested that PLMs are fascinating due to their outstanding performance, minimal need for extensive datasets during training and their versatility across different task domains. While PLMs offer significant advantages, they need to be employed carefully with suitable fine-tuning, as their effectiveness can vary depending on the specific task and dataset. Among the PLMs explored by authors, BERT and DistilBERT achieved dominant performance across several NLP tasks.

In [30], the authors have applied both machine learning classifier and deep learning classifier. Two Kaggle datasets: one of news headlines and one of Reddit reviews were utilized. CountVectorizer was applied to transform text to vectors. For modeling, the data was trained on various machine learning algorithms and then, an ensemble method was implemented by combining three classifiers: Naive

Bayes, Stochastic Gradient and Logistic Regression. Correspondingly, for deep Learning techniques, preprocessing steps such as tokenization, stop words removal, normalization were employed on the datasets. The data was split in 80:20 ratio of training and testing respectively. Evaluations metrics such as accuracy, precision, recall, f-score were utilized for performance measure. The results revealed that among other techniques, Bidirectional Encoder Representations outperformed with the accuracy of 92.73% and f-score of 93%.

The authors have focused on Arabic sarcasm detection and performed an in-depth survey on state-of-art approaches including machine learning techniques, deep Learning models and transformers-based models [32]. The outcome disclosed that classical machine learning techniques work well on small datasets. However, deep learning models which included Recurrent Neural Network (RNN) and CNN outperformed classical machine learning algorithms and were found suitable for larger datasets. Although, including large datasets will increase the processing cost and computation power. Transformers based models were further categorized into BERT-based, ELECTRA-based and GPT-based models. BERT-based models surpassed CNN and RNN and classical machine learning algorithms even with small datasets. It was concluded that transformer-based models, especially the Arabic variant of BERT outperformed the multilingual counterparts.

The authors have incorporated context-based features for sarcasm identification in [16]. A deep learning model with Bi-LSTM was employed with GloVe embedding. Likewise, the second model was built on transformer-based model BERT and feature fusion incorporating various features such as hashtag feature, sentiment related features, syntactic features and Glove embedding features. The evaluation on two twitter benchmark datasets achieved 98.5% and 98% precision respectively whereas the IAC-v2 dataset obtained 81.2% precision. The results also revealed that Bi-LSTM model could learn contextual information from sarcastic expression and using that information the model's performance can be enhanced. BERT can capture contextual information and performing feature fusion with BERT was able to handle word embedding based features.

Observations from Literature Review

From the literature review, the following observations are derived.

- Detecting sarcasm efficiently from text data is considered as one of the eminent tasks and a significant challenge in various NLP tasks.
- Recently, there has been rapid progress in the development of large language models (LLMs) across multiple application domains. Related work disclosed that LLM models deliver dominant performance on a range of natural language processing tasks, specifically, sarcasm detection.
- Indic languages - Bengali, Hindi, Gujarati, Marathi, Punjabi, Telugu and Tamil are labeled as low-resourced languages and hence, are less explored with respect to sarcasm detection.

This research aims to meet the increasing demand for precisely identifying of sarcasm in various low-resourced Indian languages. Through the generation of tailored datasets for each language and leveraging advanced pre-trained models, this research aims to support the creation of effective tools for sentiment analysis, social media monitoring and online content moderation in India's diverse linguistic environment.

The primary goal is to enhance the accuracy and efficiency of detecting sarcasm and thereby aiding in the better comprehension and management of effective communication across various linguistic contexts in India.

3 Methodology

This section describes the methodological framework to perform an empirical study on LLMs for sarcasm detection in low-resourced Indic languages. The framework is presented in Figure 1. The section gives description of data collection methods, data preprocessing and the developed strategy to perform empirical analysis.

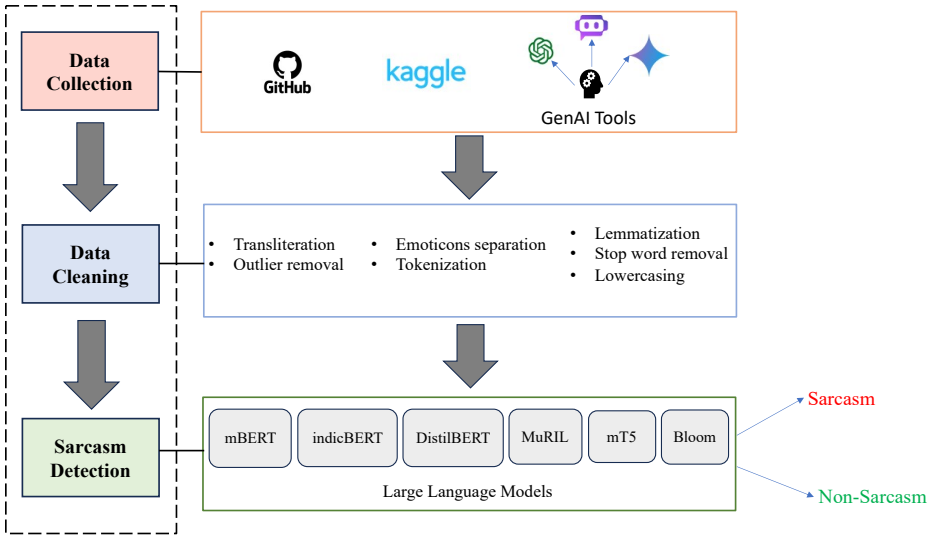

Fig. 1. Schematic Diagram of the Methodology.

3.1 Data Collection

This section describes the dataset used for our analysis. For Hindi [1,2], Bengali[3] and Telugu[4] languages, several data samples are extracted from Kaggle and Github. For Gujarati, Marathi, Punjabi and Tamil languages, synthetic data samples were generated using popular Generative Artificial Intelligence (GenAI) tools - GPT-3.5, Gemini-1.5 and PoeBot-1. Data generation is performed using a diverse range of internet text with an emphasis on conversational content. The curated datasets are meticulously crafted for each language, drawing from online content sources to ensure relevance and accuracy.

Dataset consists of various features including language, tweet dates, retweet count, tweet text, followers, followings, comment count, location, hashtags, total tweets, like count, user tag etc. Our focus is on two columns - tweet text and the label. The feature tweet text refers to the main content of the tweet and the label column depicts whether the content is sarcastic or not. Label 0 shows that the content is not sarcastic and 1 show that it is sarcastic.

Table 1 shows the data sample size for each target language. For each language, data distribution is approximately balanced. The total size of the dataset is 60,000. Figure 2 depicts the examples of sarcastic and non-sarcastic data samples from our dataset for each targeted language. Equivalent English Translation of each data sample is also provided for the understanding purpose.

[1]https://www.kaggle.com/datasets/pragyakatyayan/hindi-tweets-dataset-for-sarcasm-detection
[2]https://github.com/sbharti1984/Hindi-Tweets
[3]https://www.kaggle.com/datasets/sakibapon/banglasarc
[4]https://github.com/sbharti1984/Telugu-Sarcastic-Sentences

Table 1. Language-wise Sample Size in Dataset

Language	Hindi	Bengali	Gujarati	Marathi	Punjabi	Tamil	Telugu
Data Size (no. of samples)	8572	8572	8572	8571	8571	8571	8571
Total Sample Size for Multilingual Dataset							60,000

Language	Sample Text from Dataset	Equivalent English Translation (For Understanding)	Label
Bengali	পড়াশোনা যদি জীবন হয় তবে জীবন খুব বিরক্তিকর	If study is life, then life is very boring.	1
	বন্ধুদের আলিঙ্গন করুন, তারা মূল্যবান, প্রতিটি মুহূর্ত উদযাপন করুন	Embrace friends, they are precious, celebrate every moment	0
Gujarati	સોશિયલ મીડિયાથી દૂર જઈને, તેઓ વિશ્વને જોતા નથી, તેમનું જીવન પસંદ અને ટિપ્પણીઓમાં ઘટાડવામાં આવ્યું છે	Away from social media, they don't see the world, their lives are reduced to likes and comments	1
	ખોટા માર્ગેથી પાછા ફરવું , નિષ્ફળતા નહીં , સમજણનો પુરાવો છે	Returning from the wrong path, not failure, is evidence of understanding	0
Hindi	संगीतकार का संगीत इतना मधुर है कि आपको पागल सकता है	Musician's music is so sweet that it can drive you crazy	1
	रोज कुछ नया करो , रचनात्मक बनो जुनून बनाए रखो , प्रगति करो	Do something new every day, be creative, maintain passion, progress	0
Marathi	ही कविता वाचल्यानंतर मला वाटले की मी मुलाचे डूडल पहात आहे	After reading this poem, I felt like I was looking at a child's doodle	1
	निसर्गाच्या मांडीवर फिरा , शांतीने जा आणि इतरांना निसर्गाच्या शांतीचा धडा शिकवा	Walk in the lap of nature, walk in peace and teach others the lesson of nature's peace	0
Punjabi	ਮੈਨੂੰ ਲਗਦਾ ਹੈ ਕਿ ਮੈਂ ਅੱਜ ਜਿੰਮ ਦੇ ਤਰੀਕੇ ਨੂੰ ਭੁੱਲ ਗਿਆ ਹਾਂ	I think I forgot how to go to the gym today	1
	ਪਹਿਲੀ ਦੇ ਨੂੰ ਛੂਹਣ ਨਾਲ ਜ਼ਿੰਦਗੀ ਦੇ ਯਾਤਰਾ ਨੂੰ ਮੁਸਕਰਾਹਟ ਨਾਲ ਪ੍ਰਕਾਸ਼ਤ ਕਰੇਗਾ	Touching the first ray will illuminate life's journey with a smile	0
Tamil	இந்த வரைதல் மிகவும் மோசமானது , நான் அதை என் எதிரிக்குக் கூட காட்ட மாட்டேன்	This drawing is so bad I won't even show it to my opponent	1
	புன்னகை மிக அழகான நகைகள் , எப்போதும் அதை அணியுங்கள் , கிரிக்கவும்	Smile is the most beautiful jewelry, always wear it and smile	0
Telugu	ప్రోగ్రామింగ్ జీవితం అయితే జీవితం చాలా ఒత్తిడితో కూడుకున్నది	Programming is life but life is very stressful	1
	నది తరంగాల నుండి నేర్చుకోండి , ముందుకు సాగండి , ఎప్పుడూ ఆగవద్దు	Learn from the waves of the river, move forward, never stop	0

Fig. 2. Language-wise Samples with their Sarcasm Scores from Dataset.

3.2 Data Cleaning

As the tailored dataset contains raw data, preparation of the dataset before feeding it to the LLMs is required. The complete procedure of data preprocessing is shown in Figure 3. Total seven steps, namely, transliteration, outlier removal, emoticons separation, tokenization, lemmatization, stop word removal and lowercasing are performed during preprocessing.

Fig. 3. Data Preprocessing Steps.

During data preprocessing, firstly, the code-mixed data samples are handled. This step is called Transliteration. Code-mixed data means the actual language of the text is different from which it is written in. One simple example of transliterated text is given in Figure 4. Here, the original text is code-mixed; written in English but the actual language is Hindi. In this case, indic nlp[5] library is used for transliterating the employed dataset. The transliterated dataset is then processed to detect and remove outliers. Here, the data samples from other than the targeted languages are considered

[5]https://github.com/anoopkunchukuttan/indic_nlp_library

as outliers and removed before further processing. In tokenization and lemmatization, the tokens are generated and then converted into the respective root words. Finally, the stop words that do not contribute much to the sentiment of the content are removed and the remaining parts of the data are converted into lowercasing.

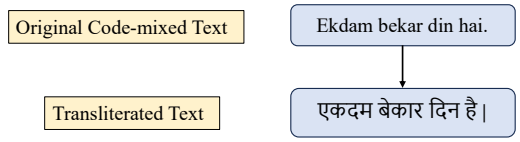

Fig. 4. An Example of Code-mixed Data.

3.3 Pre-trained Large Language Models (LLMs)

Pre-trained LLMs such as BERT, OpenAI GPT, XLNet, XLM, Bloom and MuRIL are highly utilized for NLP tasks; majorly, for two reasons: i) their state-of-the-art performance and ii) they relieve the practitioners from necessary resources to train models [34]. These models have significantly benefited both researchers and practitioners in the NLP field, especially when working with limited time and computational resources.

These pre-trained large language models are initially trained on extensive datasets without focusing on any specific task. To effectively apply them in practice, several adaptations are performed to the final output layer to align with the specific task at hand. This adaptation process is commonly known as fine-tuning. Presently, there are more than 5000 LLM models available which are utilized for either of the two purposes: i) tailored for individual task or domain or ii) optimizing the model's core functionality or reducing computation cost.

For our research, we have selected six pre-trained LLM models namely, mBERT, indicBERT, DistilBERT, MuRIL, mT5 and Bloom from literature survey. An incisive overview of these approaches is given below.

1. mBERT (multilingual Bidirectional Encoder Representations from Transformers) is renowned for its impressive performance in natural language processing tasks and comprehension [5]. Despite being a large and computationally demanding model, it remains a benchmark for various advanced applications.
2. IndicBERT is a multilingual model which incorporates 12 major Indian languages and is trained on a corpus of approximately 9 billion tokens [42]. Despite having very few parameters compared to other multilingual models, IndicBERT achieves superior performance.
3. DistilBERT is considered as a faster and lighter version of BERT [15]. DistilBERT is 60% faster than BERT. The size of the model is reduced by 40% utilizing knowledge distillation during the pre-training phase while retaining 97% of its language understanding abilities.
4. MuRIL (Multilingual Representations for Indian Languages) is trained on a substantial Indian text corpus [12]. It significantly outperforms mBERT across all tasks in the challenging XTREME benchmark.
5. mT5 is a multilingual adaptation of the T5 model pre-trained on the Common Crawl dataset with 101 languages [26]. It has extraordinary performance for various multilingual benchmarks. mT5 has significant model capacity for cross-lingual representation learning.
6. BLOOM is a large-scale language model with 176 billion parameters, developed through the collaborative efforts of hundreds of contributors and made available as an open-source resource [36]. It is trained on data of 59 languages.

3.4 Evaluation Parameters

For the evaluation of the selected LLMs, standard evaluation parameters: accuracy, precision and recall are selected based on literature survey.

Accuracy is defined as a ratio of accurately predicted results to the total results. It calculates the probability of correctness for predicted output. Accuracy is measured as shown in Equation 1. TP, TN, FP and FN are derived from the confusion matrix shown in Table 2.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

Precision is the number of true positive instances divided by the number of true positives and the number of false positives. Precision is measured as shown in Equation 2.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Recall is the number of true positive instances divided by the number of true positives and the number of false negatives. Recall is measured as shown in Equation 3.

Table 2. Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

4 Result Analysis

The result analysis is performed in three phases. In the first phase, the performance of the considered six LLMs: mBERT, indicBERT, DistilBERT, MuRIL, mT5 and Bloom is analyzed across the selected seven languages: Bengali, Hindi, Gujarati, Marathi, Punjabi, Telugu and Tamil. In the second phase, for each of the targeted language, the performance of the considered six LLMs is compared. In the third phase, the performance of six LLMs are compares on the multilingual dataset which consists all the data samples of each of seven considered languages.

Figure 5 depicts the phase 1 result analysis. It is observed that mBERT achieved high accuracy rates across Bengali (98.3%), Gujarati (99.2%), Hindi (99.89%), Marathi (98.9%), Punjabi (96.4%), Tamil (99.78%) and Telugu (99.5%). Similarly, indicBERT demonstrated strong performance, particularly in Tamil (99.67%), Punjabi (99.1%), and Gujarati (98.4%). DistilBERT showcased notable accuracy in Marathi (99.2%), Telugu (98.86%) and Hindi (98.52%). Muril exhibited high accuracy in languages like Marathi (99.9%) and Hindi (99.89%). Meanwhile, mT5 and Bloom models displayed varied performance across different languages, with mT5 achieving significant accuracy in Punjabi (99.4%) and Gujarati (99.2%). Overall, these metrics provide insights into the effectiveness of each model in accurately predicting sarcasm across diverse linguistic contexts.

Figure 6 demonstrate the result analysis of phase 2. It is observed that the performance of different LLM varies across diverse languages. Among other LLMS, MuRIL gives average accuracy of 99% across all the seven languages. For Bengali, Gujarati, Hindi, Marathi, Tamil and Telugu; MuRIL and mBERT achieved high accuracy among other models whereas for Punjabi, mT5 and MuRIL outperforms other models. Overall, MuRIL, mBERT, mT5, indicBERT, DistilBERT and Bloom give average performance (in the same order) across all seven languages.

Figure 7 shows the accuracy of training process for all the six LLMs. The model training is performed for ten epochs. It has been observed that mBERT demonstrates significant learning efficiency as the number of epochs increases, while other models show a steady improvement with more epochs. MuRIL exhibited increasingly efficient learning accuracy with each successive epoch.

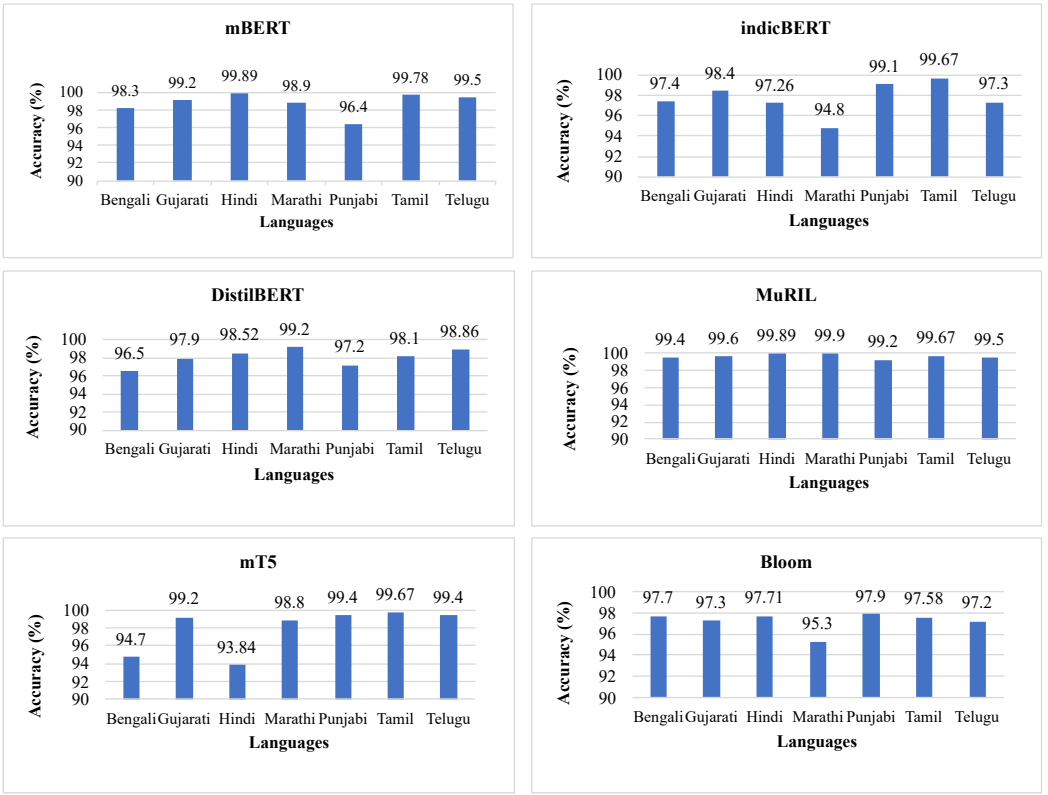

Fig. 5. Performance Comparison of each Considered LLM Models across Diverse Languages.

Figure 8 presents the results analysis for phase 3. For the curated multi-lingual dataset, the performance of all the models is compared with respect to accuracy, precision and recall. From the results, it is observed that MuRIL outperforms other models in terms of all the parameters for multilingual dataset. MuRIL is specially designed to manage multilingual representations for Indian languages and is trained on a large corpus of Indian text, enabling it to achieve superior performance. After MuRIL, the variants of BERT model: mBERT, DistilBERT and indicBERT perform efficiently.

5 Conclusion and Future Directions

This research concentrates on tackling the significant challenge of sarcasm in natural language processing (NLP) systems. Sarcasm is a linguistic phenomenon characterized by its nuanced and context-sensitive characteristics. Specifically targeting Bengali, Hindi, Gujarati, Marathi, Punjabi, Telugu and Tamil languages, the research endeavors to devise a comprehensive solution for detecting sarcasm within these languages. Recognizing the diverse linguistic landscape of India, the research aims to develop custom datasets for each language and harness advanced pre-trained models. Customized datasets have been carefully created for each language, utilizing online content sources to guarantee both relevance and precision. The methodology presented in this research leverages state-of-the-art pre-trained large language models to facilitate the training and evaluation of the sarcasm detection system.

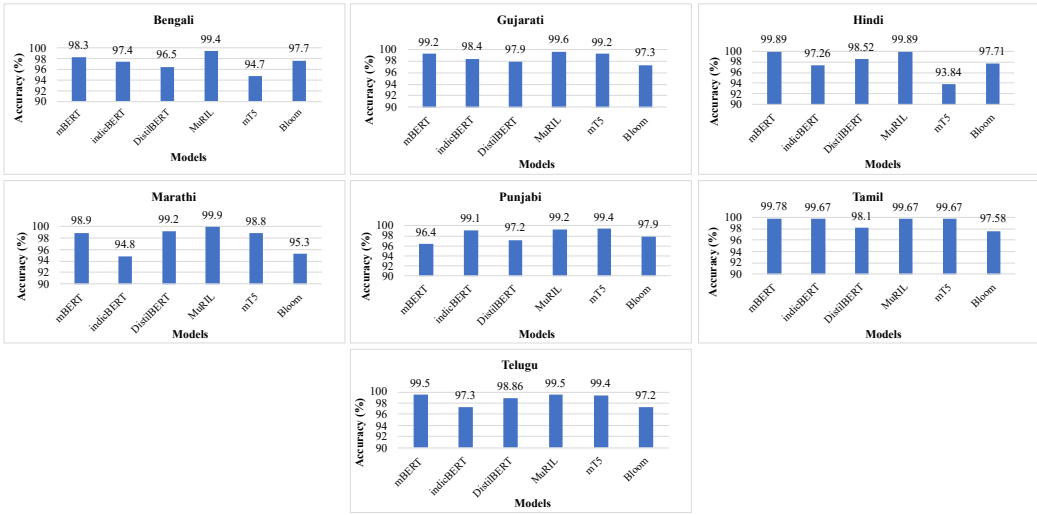

Fig. 6. Performance Comparison of Diverse LLM Models for each Considered Language.

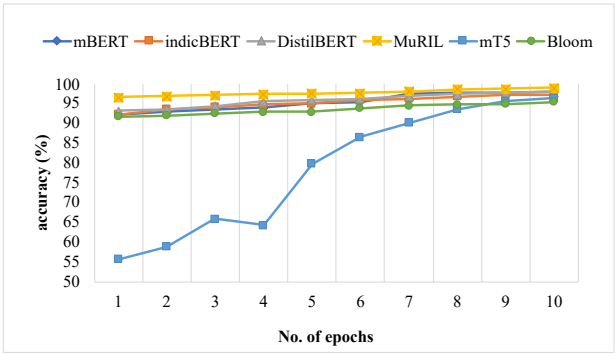

Fig. 7. Epoch-wise Training Accuracy of Diverse LLM Models for Multilingual Dataset.

The outcome of this research will help in enhancing the effectiveness and adaptability of natural language processing systems in recognizing and interpreting sarcasm across diverse linguistic contexts. This will aid in developing effective tools for sentiment analysis, social media monitoring and online content moderation. Ultimately, this will improve the effectiveness and accuracy of sarcasm detection, leading to a better understanding and management of online interactions across various linguistic contexts within the country.

References

[1] Malak Abdullah, Jumana Khrais, and Safa Swedat. 2022. Transformer-based deep learning for sarcasm detection with imbalanced dataset: Resampling techniques with downsampling and augmentation. In *2022 13th International Conference on Information and Communication Systems (ICICS)*. IEEE, 294–300.

[2] Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F Bissyandé, Jacques Klein, and Anne Goujon. 2021. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*. 260–268.

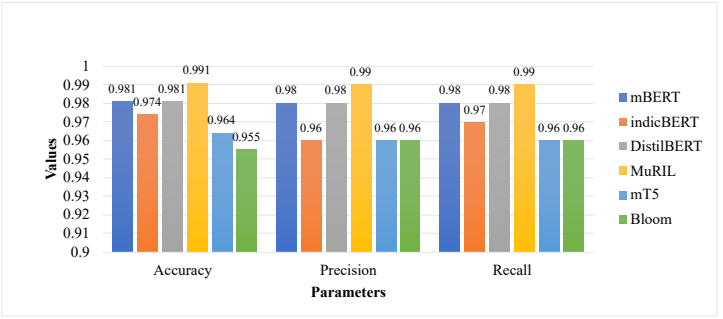

Fig. 8. Performance Comparison of Diverse LLM Models for Multilingual Dataset.

[3] Ishana Attri and Maitreyee Dutta. 2019. Bi-Lingual (English, Punjabi) Sarcastic Sentiment Analysis by using Classification Methods. *IEEE Access* (2019), 1385.

[4] Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N Govaichelvan. 2023. Tamil offensive language detection: Supervised versus unsupervised learning approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 4 (2023), 1–14.

[5] David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, Vol. 9. 574–577.

[6] Vibhuti Bansal, Mrinal Tyagi, Rajesh Sharma, Vedika Gupta, and Qin Xin. 2022. A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages. *ACM transactions on Asian and low-resource language information processing* (2022).

[7] Santosh Kumar Bharti, Korra Sathya Babu, and Rahul Raman. 2017. Context-based sarcasm detection in Hindi tweets. In *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 1–6.

[8] Santosh Kumar Bharti, Rajeev Kumar Gupta, Prashant Kumar Shukla, Wesam Atef Hatamleh, Hussam Tarazi, and Stephen Jeswinde Nuagah. 2022. Multimodal sarcasm detection: a deep learning approach. *Wireless Communications and Mobile Computing* 2022, 1 (2022), 1653696.

[9] Santosh Kumar Bharti, Reddy Naidu, and Korra Sathya Babu. 2020. Hyperbolic feature-based sarcasm detection in Telugu conversation sentences. *Journal of Intelligent Systems* 30, 1 (2020), 73–89.

[10] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2017. Harnessing online news for sarcasm detection in hindi tweets. In *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 679–686.

[11] Santosh Kumar Bharti, Bakhtyar Vachha, RK Pradhan, Korra Sathya Babu, and Sanjay Kumar Jena. 2016. Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks* 2, 3 (2016), 108–121.

[12] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems* 32 (2019).

[13] Mitali Desai, Rupa G Mehta, and Dipti P Rana. 2019. An empirical analysis to identify the effect of indexing on influence detection using graph databases. *International Journal of Innovative Technology and Exploring Engineering* 8, 9S (2019), 414–21.

[14] Mitali Desai, Rupa G Mehta, and Dipti P Rana. 2024. Anatomising the impact of ResearchGate followers and followings on influence identification. *Journal of Information Science* 50, 3 (2024), 607–624.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Christopher Ifeanyi Eke, Azah Anir Norman, and Liyana Shuib. 2021. Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. *IEEE Access* 9 (2021), 48501–48518.

[17] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. 2020. Sarcasm identification in textual data: systematic review, research challenges and open directions. *Artificial Intelligence Review* 53 (2020), 4215–4258.

[18] Priya Goel, Rachna Jain, Anand Nayyar, Shruti Singhal, and Muskan Srivastava. 2022. Sarcasm detection using deep learning and ensemble learning. *Multimedia Tools and Applications* 81, 30 (2022), 43229–43252.

[19] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 1–22.

[20] Kanhaiyya Khandagale and Hetal Gandhi. 2022. Sarcasm Detection in Hindi-English Code-Mixed Tweets Using Machine Learning Algorithms. In *International Conference on Computing in Engineering & Technology*. Springer, 221–229.

[21] Akshi Kumar and Geetanjali Garg. 2023. Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets. *Journal of ambient intelligence and humanized computing* 14, 5 (2023), 5327–5342.

[22] Akshi Kumar, Saurabh Raj Sangwan, Adarsh Kumar Singh, and Gandharv Wadhwa. 2023. Hybrid deep learning model for sarcasm detection in Indian indigenous language using word-emoji embeddings. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 5 (2023), 1–20.

[23] R Prasanna Kumar, G Bharathi Mohan, Yamani Kakarla, SL Jayaprakash, Kolla Gnapika Sindhu, Tekumudi Vivek Sai Surya Chaitanya, Bachu Ganesh, and Nunna Hasmitha Krishna. 2023. Sarcasm detection in Telugu and Tamil: an exploration of machine learning and deep neural networks. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 1–7.

[24] Amit Kumar Bhadra, SG Shaila, and MK Banga. 2022. Review on sentiment analysis and polarity classification of sarcastic sentences using deep learning in social media. *Data Engineering and Intelligent Computing: Proceedings of 5th ICICC 2021, Volume 1* (2022), 225–237.

[25] Florian Kunneman, Christine Liebrecht, Margot Van Mulken, and Antal Van den Bosch. 2015. Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management* 51, 4 (2015), 500–509.

[26] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. (2023).

[27] Sanzana Karim Lora, GM Shahariar, Tamanna Nazmin, Noor Nafeur Rahman, Rafsan Rahman, Miyad Bhuiyan, et al. 2022. Ben-sarc: A corpus for sarcasm detection from bengali social media comments and its baseline evaluation. (2022).

[28] Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open* 4 (2023), 13–18.

[29] Moumita Pal and Rajesh Prasad. 2023. Sarcasm detection followed by sentiment analysis for Bengali language: neural network & supervised approach. In *2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS)*. IEEE, 1–7.

[30] Ameya Parkar and Rajni Bhalla. 2024. Analytical comparison on detection of Sarcasm using machine learning and deep learning techniques. *International Journal of Computing and Digital Systems* 15, 1 (2024), 1615–1625.

[31] Pravin K Patil and Satish R Kolhe. 2023. Sarcasm Detection for Marathi and the role of emoticons. In *International Conference on Data Intelligence and Cognitive Informatics*. Springer, 193–204.

[32] Alaa Rahma, Shahira Shaaban Azab, and Ammar Mohammed. 2023. A comprehensive survey on Arabic sarcasm detection: approaches, challenges and future trends. *IEEE Access* 11 (2023), 18261–18280.

[33] Ratnavel Rajalakshmi, M Saptharishee, S Hareesh, R Gabriel, et al. 2024. DLRG-DravidianLangTech@ EACL2024: Combating Hate Speech in Telugu Code-mixed Text on Social Media. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. 140–145.

[34] Md Saifullah Razali, Alfian Abdul Halin, Lei Ye, Shyamala Doraisamy, and Noris Mohd Norowi. 2021. Sarcasm detection using deep learning with contextual features. *IEEE Access* 9 (2021), 68609–68618.

[35] Daniel Šandor and Marina Bagić Babac. 2023. Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery* ahead-of-print (2023).

[36] Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8766–8774.

[37] Bhumi Shah and Margil Shah. 2021. A survey on machine learning and deep learning based approaches for sarcasm identification in social media. In *Data Science and Intelligent Applications: Proceedings of ICDSIA 2020*. Springer, 247–259.

[38] D Sumathi, B Gowtham, K Naveen, and H Subramani. 2021. Sentiment classification on Tamil and Telugu text using RNNs and Transformers. In *2021 International Conference on Technological Advancements and Innovations (ICTAI)*. IEEE, 582–587.

[39] Madhuri Thorat and Nuzhat Faiz Shaikh. 2024. Novel Deep Neural Network Approach for the Sarcasm Detection in Hindi Language. *International Journal of Intelligent Systems and Applications in Engineering* 12, 10s (Jan. 2024), 487–494.

[40] David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2023. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing* 14, 6 (2023), 7399–7410.

[41] Oxana Vitman, Yevhen Kostiuk, Grigori Sidorov, and Alexander Gelbukh. 2023. Sarcasm detection framework using context, emotion and sentiment features. *Expert Systems with Applications* 234 (2023), 121068.

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).