

SARCASM AND MENTAL HEALTH DETECTION FROM SOCIAL MEDIA USING TRANSFORMER-BASED TECHNIQUES

Subject Code RIRA602

**Thesis Submitted
in Partial Fulfillment of the Requirements for the Degree of**

**MASTER OF TECHNOLOGY
BY RESEARCH**

in

COMPUTER SCIENCE AND ENGINEERING

by

SUMAN SAHA

(24/RCO/03)

**Under the Supervision of
Dr. SANJAY KUMAR
Department of Computer Science and Engineering, Delhi
Technological University**



Department of Computer Science and Engineering

**DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042**

May, 2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to **Dr. SANJAY KUMAR** for his continuous guidance and mentorship that he provided during research work. He showed me the path to achieving targets by explaining all the tasks to be done and explained to me the importance of this work as well as its industrial relevance. He was always ready to help me and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this work would not have been successful.

Place: New Delhi

Date: 27/05/2026

Suman Saha
SUMAN SAHA

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **SUMAN SAHA (24/RCO/03)**, of M.Tech By Research (CSE), hereby certify that the work which is being presented in the thesis entitled **“Sarcasm and Mental Health Detection from Social Media Using Transformer-Based Techniques”** in partial fulfillment of the requirement for the award of the degree of Master of Technology By Research, submitted in the Department of Computer Science and Engineering, Delhi Technological University, is an authentic record of my own work carried out under the supervision of **Dr. SANJAY KUMAR**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.



Signature of Supervisor



Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that **SUMAN SAHA (24/RCO/03)** has carried out the research work presented in this thesis entitled “**Sarcasm and Mental Health Detection from Social Media Using Transformer-Based Techniques**” for the award of Master of Technology By Research from the Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.



(Dr. SANJAY KUMAR)

Assistant Professor

Department of Computer Science and Engineering

Delhi Technological University

Date: 27/05/2026

SARCASM AND MENTAL HEALTH DETECTION FROM SOCIAL MEDIA USING TRANSFORMER-BASED TECHNIQUES

SUMAN SAHA

ABSTRACT

Social media platforms generate massive amounts of textual content every day, containing valuable information related to human emotions, opinions, sarcasm, and psychological behavior. Understanding such contextual and emotionally implicit information has become an important research problem in Natural Language Processing and Artificial Intelligence. This thesis presents transformer-based approaches for sarcasm detection and mental health classification from social media text using contextual representation learning and Explainable Artificial Intelligence techniques. For sarcasm detection, a transformer-based framework named SarcasmBERTa is proposed by integrating RoBERTa with Multi-Head Attention mechanisms to capture contextual and semantic relationships from multilingual social media text. The framework was evaluated on English, Hindi, and Bengali datasets and demonstrated superior performance compared to several baseline and contemporary models. For mental health detection, the proposed X-MHD framework combines DistilBERT, knowledge graph embeddings, and Explainable Artificial Intelligence techniques. The model effectively identifies emotional well-being conditions such as depression, anxiety, bipolar disorder, and suicide-related behaviour from online textual content.

LIST OF PUBLICATIONS

- Suman Saha and Sanjay Kumar, “ SarcasmBERTa: Multilingual Sarcasm Detection using RoBERTa and Multi-Head Attention Based Model,” accepted with minor revisions to ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2026.
- Suman Saha and Sanjay Kumar, “ Explainable Multi-Class Mental Health Classification from Social Media Using DistilBERT and Knowledge Graphs,” 2026 Fourth International Conference on Secure Cyber Computing and Communication (ICSCCC 2026), Accepted.

Contents

LIST OF PUBLICATIONS	1
LIST OF SYMBOLS	6
1 INTRODUCTION	7
1.1 Problem Statement and Objectives	7
1.1.1 Sarcasm Detection Using Transformer-Based Techniques	8
1.1.2 Mental Health Detection Using Explainable Transformer Framework	8
1.2 Motivation	8
2 LITERATURE REVIEW	10
2.1 Baseline and Deep Learning-Based Models	10
2.2 Transformer and Pretrained Language Model (PLM)-Based Models	11
3 METHODOLOGY	16
3.1 Overview of Proposed Framework	16
3.1.1 Sarcasm Detection Framework	16
3.1.2 Mental Health Detection Framework	25
4 RESULT & ANALYSIS	30
4.1 Results for Sarcasm Detection	30
4.1.1 Hyperparameter Study	31
4.1.2 Performance comparison with Baseline Algorithms and Contemporary Models	31
4.1.3 Evaluation on Bengali Datasets	32
4.1.4 Evaluation on Hindi Datasets	33
4.1.5 Ablation Study	37
4.2 Results for Mental Health Detection	37
4.2.1 Ablation Study	37

5	CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT	44
5.1	Conclusion	44
5.2	Future Scope and Social Impact	45
	References	53
	Proof of Publishing	
	Accepted / Published Papers	
	Plagiarism Report	

List of Tables

4.1	Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Dataset 1 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).	32
4.2	Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Dataset 2 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).	33
4.3	Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Datasets 3, 4, and 5 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).	41
4.4	An ablation study to evaluate the impact of the attention mechanism used in the SarcasmBERTa model.	42
4.5	Ablation Study of X-MHD	42
4.6	Summary of Experimental Results on the Test Dataset	42
4.7	Per-Class Performance of X-MHD	43

List of Figures

3.1	Architecture of the proposed sarcasm detection SarcasmBERTa model	19
3.2	Workflow of the Sarcasm Detection Model, i.e., SarcasmBERTa	20
3.3	Overall architecture of the proposed X-MHD model integrating DistilBERT, Knowledge Graph features, and LIME-based explainable AI for multi-class mental health disorder detection.	28
4.1	Representation of Training and Validation Loss over Epochs .	35
4.2	Representation of Training and Validation Accuracy over Epochs	35
4.3	Confusion Matrix for Sarcasm Detection Model	36
4.4	F1-score comparison across mental health conditions.	38
4.5	LIME-Based Local Explanations for Token-Level Contributions to Predicted Mental Health Category.	39
4.6	LIME-Based Local Explanation for Suicide Watch Prediction, with Token-Level Contributions.	40

LIST OF SYMBOLS

AI	Artificial Intelligence
NLP	Natural Language Processing
XAI	Explainable Artificial Intelligence
LLM	Large Language Model
MHA	Multi-Head Attention
KG	Knowledge Graph
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
DistilBERT	Distilled Version of BERT
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
SVM	Support Vector Machine
Acc	Accuracy
Pre	Precision
R	Recall
F1	F1-Score
CL	Consistency Loss
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
Q	Query Matrix
K	Key Matrix
V	Value Matrix
H	Hidden Representation Matrix
$\mathbb{R}$	Set of Real Numbers
$\mathbb{Z}$	Set of Integer Numbers

Chapter 1

INTRODUCTION

The rapid growth of social media platforms such as Twitter, Reddit, Facebook, and Instagram has transformed the way individuals communicate, express opinions, and share emotions in daily life. Millions of users continuously post thoughts, reactions, experiences, and personal feelings on online platforms, generating massive amounts of textual data every day. This user-generated content contains valuable information related to human emotions, behaviour, sentiment, and psychological conditions. In recent years, researchers have increasingly focused on using NLP and AI techniques to analyze social media text for applications such as sentiment analysis, sarcasm detection, emotion recognition, and mental health assessment. The informal and context-dependent nature of social media conversations makes understanding textual meaning a challenging task, particularly when users express emotions through sarcasm, humour, irony, or indirect language. Therefore, developing intelligent and interpretable transformer-based systems for understanding social media text has become an important research area in modern NLP and healthcare-oriented AI applications.

1.1 Problem Statement and Objectives

Social media users frequently express emotions, stress, opinions, and psychological conditions through indirect language, sarcasm, humour, and implicit contextual expressions. Detecting such patterns accurately remains a challenging task due to contextual ambiguity, multilingual variations, and lack of semantic understanding in traditional NLP systems. Existing methods also suffer from poor interpretability and limited capability in understanding low-resource languages.

This thesis addresses these limitations by proposing transformer-based frameworks for sarcasm and mental health detection from social media text using contextual representation learning, attention mechanisms, and Explainable Artificial Intelligence techniques.

1.1.1 Sarcasm Detection Using Transformer-Based Techniques

Sarcasm detection is one of the most challenging problems in Natural Language Processing because sarcastic statements often express meanings opposite to their literal interpretation. Traditional sentiment analysis systems fail to correctly identify such hidden emotional intentions, leading to inaccurate sentiment prediction and contextual understanding.

The objective of this work is to develop a multilingual transformer-based sarcasm detection framework using pretrained language models such as RoBERTa integrated with Multi-Head Attention mechanisms. The proposed approach focuses on improving contextual understanding, semantic feature representation, and sarcasm classification performance across English, Hindi, and Bengali datasets.

1.1.2 Mental Health Detection Using Explainable Transformer Framework

Mental health disorders such as depression, anxiety, bipolar disorder, and suicidal ideation are increasingly reflected in social media conversations. Early identification of such psychological conditions through automated NLP systems can assist researchers and healthcare professionals in understanding emotional well-being and behavioural patterns. The objective of this research is to develop an explainable transformer-based mental health detection framework using DistilBERT, semantic knowledge integration, and Explainable Artificial Intelligence techniques such as LIME and attention visualization. The proposed framework aims to improve classification accuracy, interpretability, and robustness for detecting mental health conditions from social media text.

1.2 Motivation

Increase in expansion of social media platforms has significantly increased the amount of user-generated textual data containing emotions, opinions, sarcasm, humour, and psychological expressions. Platforms such as Twitter, Reddit, and Facebook have become important spaces where individuals openly express their thoughts, frustrations, emotional stress, and mental

health conditions. This growing availability of textual data has motivated researchers to explore advanced Natural Language Processing (NLP) techniques for understanding human behaviour and emotional patterns from on-line communication. One of the major motivations behind this research is the limitation of traditional sentiment analysis and machine learning approaches in understanding contextual and implicit meanings present in social media text. Sarcasm often conveys emotions opposite to the literal meaning of a sentence, making sentiment understanding highly challenging. Incorrect interpretation of sarcastic statements can negatively affect emotion analysis and downstream applications such as mental health prediction systems. Many individuals express their psychological conditions indirectly through sarcastic or emotionally complex statements on social media platforms. Therefore, accurate detection of sarcasm and contextual understanding becomes essential for improving mental health analysis systems. BERT, RoBERTa models provide the capability to capture long-range contextual dependencies, implicit meanings, and emotional cues more effectively than traditional deep learning approaches. This motivated the use of transformer-based frameworks for both sarcasm detection and mental health classification tasks. Furthermore, most existing systems operate as black-box models and lack interpretability, which reduces trustworthiness in healthcare-related applications. To provide transparent and interpretable predictions, XAI techniques such as LIME is used.

The motivation of this thesis is therefore to develop an efficient, multilingual, and explainable transformer-based framework capable of accurately detecting sarcasm and mental health conditions from social media text while improving contextual understanding, robustness, and prediction interpretability.

Chapter 2

LITERATURE REVIEW

The rapid advancement of NLP and AI has improved the analysis of social media text for applications such as sentiment analysis, sarcasm detection, emotion recognition, and mental health assessment. Social media platforms generate large volumes of user-generated textual data containing opinions, emotions, humour, and psychological expressions. Researchers have increasingly explored machine learning, deep learning, and transformer-based approaches to understand contextual meaning and emotional behaviour from such data.

2.1 Baseline and Deep Learning-Based Models

Ahuja and Sharma [2] propose the LMTweets encoder model for sarcasm and irony detection, trained on 500,000 tweets. Cambria et al. [35] proposed a multi-task learning framework based on a deep neural network to model the connection between sentiment and sarcasm while simultaneously solving both tasks. Cai et al. [9] developed a multi-modal sarcasm detection model for Twitter that combined text, images, and image attributes. Their hierarchical fusion model first extracted features from each modality and then integrated them using a bidirectional LSTM network. Zhang et al. [57] created ConAttSD, a model designed to detect sarcasm in multimodal situations, like video conversations. Use of a contrastive attention mechanism, which highlights the differences between positive text paired with a negative tone. Collection of 1.3 million sarcastic statements from Reddit that were self-annotated, which was the largest sarcasm dataset done by Khodak et al. [27]. Rosid et al. [46] developed a hybrid model called MHA-CovBi for detecting sarcasm in Indonesian-English mixed-code texts. The model combined CNN with multi-head attention and BiGRU, using Fast-Text for Indonesian and GloVe for English word embeddings. Pandey and Singh [40] proposed a sarcasm detection model by combining LSTM and BERT for code-mixed language. Yue et al. [56] developed KnowleNet, a

multimodal sarcasm detection model that used ConceptNet to incorporate prior knowledge and improve understanding of image-text by detecting cross-modal semantic similarities. A new hybrid model was created by Hazarika et al. [19] that tried to detect sarcasm by analyzing authors’ past behavior in comments and studying the general behavior of authors within specific subreddits. Bao et al. [5] introduced the HIAN model to address the challenge of aligning diverse sarcasm cues from text and images. It utilized a BiLSTM with interactive attention to extracting richer textual features by integrating visual class words. Application of machine learning (ML) techniques for analyzing mental health has been widely investigated in previous studies. Early research work mainly focused on traditional techniques of ML models to detect emotional and psychological patterns in textual data. Kaushik et al. [25] studied classical algorithms, including SVM, also Logistic Regression, and Multinomial Naïve Bayes, for sentiment analysis on mental health datasets. Focusing on early detection and intervention, D’Cruz et al. [16] developed ML classifiers based on some contexts gathered from Twitter and Reddit to predict depressive tendencies based on user-generated posts. Similarly, Chen et al. [13] introduced the psychological effects of the COVID-19 pandemic through textual analysis of comments on the website Reddit that were sourced from mental-health related datasets, such as anxiety, depression, SuicideWatch, extending this line of research to multilingual environments, Dwivedi and Rao [15] introduced a hybrid architecture that combines Recurrent Neural Networks (RNNs) with Bidirectional LSTM (Bi-LSTM) models for identifying mental health-related signals in multiple languages. Sharma et al. [49] proposed the ML methods for analyzing mental fitness issues in workplace environments. A significant breakthrough in natural language processing began with the advent of the architecture based on transformer proposed by Vaswani et al [52]. This paradigm shift marked the beginning of influential pretrained text models, become central to mental health and sentiment analysis since.

2.2 Transformer and Pretrained Language Model (PLM)-Based Models

A new approach for sarcasm detection that used context from parent comments on Reddit is proposed by Yang et al. [55]. They worked with the

SARC dataset and employed a pre-trained XLNet model to extract features. An approach for detecting sarcasm in Twitter posts using the transformer model has been proposed by Gregory et al. in [18]. The authors also used LSTM, GRU, and transformer models. The top-performing model in this approach is an ensemble of transformer models, including BERT, RoBERTa [32], XLNet [55], RoBERTa-large, and ALBERT [30]. At some points in the text, an emoji is used, followed by how it is interpreted by the model, and then how it is used in a model which incorporates it SEEmoji MUSTARD [12], an extended dataset which includes emojis and uses them for sentiment and emotion annotations for better sarcasm detection in multimodal conversations. MUSTARD dataset [11] is utilized for multimodal-based sarcasm detection. The dataset is based on TV shows, which consist of audio-visual dialogues with sarcasm labels and the dialogue history. Internet community sites such as Reddit and Twitter are highly populated with figurative language. Nikhil Jaiswal [24] suggested a deep neural structure for the detection of sarcasm. He utilized pre-trained language models such as BERT and RoBERTa. He then used them to improve their performance on a Twitter dataset. A sarcasm detection model is proposed by Meng et al. [37], combining BERT and an attention mechanism to address the challenge of detecting sarcasm in sentences with complex semantics and few sentiment words. However, sarcasm’s implicit nature makes it challenging to detect, [3], particularly in Bangla, where research remains limited compared to English. Anan et al. introduced a BERT-based model, which was much better than traditional data-driven techniques. To test their approach, they created BanglaSarc, a new Bangla sarcasm dataset made up of comments from Facebook and YouTube. HumourHindiNet, a framework for detecting humor in Hindi web series, was created by Kumar et al. [28]. They built a new multimodal dataset using dialogues from shows like ‘Kota-Factory’ and ‘Panchayat’, which were manually labeled as humor or non-humor. Pramanick et al. [42] introduced a new framework called MuLOT (Multimodal Learning using Optimal Transport). This framework has been tested using benchmark datasets such as MUSTARD, UR-FUNNY, and MST. [10]. Combining textual linguistic features with visual information, the multimodal sarcasm detection approach was proposed by Han et al. [9]. Their model focused on the textual modality, extracting high-frequency characters from pure text and hashtags to create a third modality alongside images and text. Qin et al. [43] created MMSD 2.0, an improved version of the original MMSD dataset, to fix problems like misleading cues and unrealistic negative sample. Some-

times language comes into the problem. Bedi et al. [6] developed MaSaC, a Hindi-English code-mixed dataset for sarcasm detection and humor classification in conversations, and proposed MSH-COMICS, an attention-rich neural model. Liang et al. [31] introduced an approach to address the challenge of identifying sarcasm using both text and images. The method employed object detection and OCR to extract key visual and textual cues, which were modeled using multimodal graphs. A graph-based contrastive learning strategy was proposed to enhance modality interaction and representation. In multimodal, we need to combine text, images, and image attributes. A huge corpus of Bengali text for the specific problem of research on sarcasm detection in Bengali, named ‘Ben-Sarc,’ comprising a total of 25,636 comments gathered manually from various Facebook public pages and rated by external evaluators, was proposed by Lora et al. [33]. Pokhriyal and Jain [41] proposed a sarcasm detection approach for Hindi social media comments using a probabilistic non-linear optimization technique integrated with word embeddings. For the Algerian dialect, Bousmaha et al. [8] introduced an integrated multimodal approach to sarcasm detection by utilizing both visual cues through VGG-19 and text cues through EasyOCR with some linguistic properties and machine learning algorithms. Zhang et al. [58] proposed a Multimodal Interaction Learning (MIL) framework for joint sarcasm and sentiment detection that employs cross-modal target attention and a dual-gating network to capture shared and task-specific features across modalities. Rodríguez et al. [45] proposed SINCERE, a hybrid graph-based sarcasm detection framework that integrates sentiment and emotion analysis using distant supervision and graph-based latent representation learning. Zhang et al. [59] introduced AMSD Dataset and the SarcasmBench benchmark to evaluate sarcasm understanding across 16 LLMs. While previous models performed well on certain datasets, they still had limitations in multilingual and low-resource settings. By addressing this issue we proposed a model that overcame these limitations and achieved good performance. Refai et al. [44] proposed a multi-dimensional prompt evaluation framework for LLMs that assesses prompts based on similarity, performance, efficiency, and consistency across several Arabic NLP tasks—including sarcasm detection—using models such as GPT-4o mini, LLaMA, and Claude 3.5 Haiku, highlighting trade-offs between prompt effectiveness and computational efficiency. Bade et al. [4] investigated pragmatic generalization of LLMs for multilingual sarcasm detection by fine-tuning encoder models such as BERT, RoBERTa, XLM-RoBERTa, and DistilBERT, while comparing them with few-shot GPT-4o,

showing that fine-tuned models significantly outperform generative LLMs in cross-lingual sarcasm detection tasks. According to Ou & Li et al. [39], a multi-granularity information fusion network can be used for the detection of multimodal sarcasm using multi-task CLIP and global-local cross-modal interaction learning. Various machine learning and deep learning architectures, including Random Forest, XGBClassifier, fastText-stackedTCN, and mBERT, were explored by Ahmad et al. [1] on a dataset of 1,945 humorous Hindi newspaper headlines from Hindustan and Dainik Jagran. Although previous models achieved success on certain datasets, they are still not sufficient for low-resource and multilingual scenarios. We proposed a model that transcends these limitations and achieves success on these problems. Bikku et al. [7] assessed the performance of BERT on large-scale sentiment analysis on social media data. In a comparative study, Joshy et al. [23] assessed in mental health domain using BERT, DistilBERT, RoBERTa. Ji et al. [22] introduced two model MentalBERT and MentalRoBERTa, which were pretrained on Reddit data from mental health communities. In a related effort, Vajre et al. [51] proposed PsychBERT, a model pretrained on a combination of biomedical literature and mental health-focused social media data. Madhuri et al. [34] introduced a mental health screening approach for homeless children by analyzing responses to the Pediatric Symptom Checklist (PSC) by using a SVM model. Kumar et al. [29] proposed a smartwatch-based deep learning framework for mental stress detection that addresses the limitation of missing respiration rate in wearable sensors. Sarode et al. proposed [48] a decision support model to assess Mental Health Literacy (MHL) by segmenting individuals in academic communities using modifiable predictors. Addressing the difficulties related to data scarcity also the lack of open-source mental health-focused LLMs. MentaLLaMA, an open-source LLM family built upon the LLaMA2 architecture is proposed by Yang et al. [54]. Yadav et al. [47] introduced an AI-powered mental wellness chatbot that leveraged generative AI and embedding-based relevance detection to deliver context-aware and empathetic responses. Marrapu et al. [36] proposed a data science-based model for predicting public mental health conditions using sensor data and healthcare datasets. Existing studies rarely combine contextual language modeling, structured knowledge integration, and explainable AI within a unified framework. Taking inspiration from it, our paper introduces a novel hybrid architecture, named X-MHD, which incorporates DistilBERT representations, knowledge graph embeddings, and explainable artificial intelligence to achieve interpretable mental health clas-

sification which consists multiple classes using text taken from social media. The next section overviews the architecture and components of the proposed X-MHD approach.

Chapter 3

METHODOLOGY

3.1 Overview of Proposed Framework

The proposed framework integrates contextual language understanding, semantic feature representation, attention mechanisms, and interpretability to improve classification performance and prediction transparency. The proposed framework consists of two major components: multilingual sarcasm detection and explainable mental health detection. For sarcasm detection, a pretrained RoBERTa model integrated with Multi-Head Attention (MHA) mechanisms is utilized to capture implicit semantic relationships and contextual cues present in sarcastic expressions. For mental health detection, DistilBERT along with knowledge graph embeddings is employed to enhance semantic understanding of psychological concepts and emotional patterns. Sarcasm detection and mental health analysis are closely connected research problems because users often express emotional distress, stress, anxiety, or frustration through sarcastic and indirect language. Incorrect interpretation of sarcastic statements may negatively affect mental health prediction systems. Therefore, contextual understanding of sarcasm plays an important role in improving emotional and psychological analysis from social media conversations.

3.1.1 Sarcasm Detection Framework

Data Processing

We start with data that is already public, although it is not clean. It is raw and contains clutter like URLs and image IDs. The data used in this work is from KnowleNet, which needs to be cleaned. For ‘Ben-Sarc,’ the labels are already provided as “sarcasm” and “not sarcasm.” For our work, we standardize the labels by using 1 for “sarcasm” and 0 for “not sarcasm.” We use the same labels for all the data. While cleaning the data, we realized that there are duplicate data in some of the data sets, such as ‘Ben-Sarc’ and Hindi. For ease, we also clean the paragraph formatting in the Hindi data. For k-fold validation, we use all the validation, test, and training data in one

dataset for each language, such as English, Hindi, and Bengali. Our method helps in making the process of preprocessing and validation in various data sets efficient.

Tokenization: Preprocessing Text for Model Input

Roberta tokenizer is used in this paper. This tokenizer is important in facilitating the transformation of the unstructured input into an understandable language for the model. This tokenizer is initialized with the weight parameters of the ‘roberta-base’ model from Hugging Face. To ensure standardization in input data, truncation, and padding were used where necessary.

- Truncation: Inputs with more than 256 tokens were truncated to the defined max limit, emphasizing key content.
- Padding: Tokens were used for shorter input texts to ensure consistency in data format.

For each input text, tokenization produces two outputs: attention masks and token ID outputs. The model is capable of processing text data in numerical form, token ID, which is an integer value allocated to each token. The attention masks represent valid token positions (1) and padding positions (0). The tokenizer processes inputs in batches, thereby minimizing time delay in data flow.

Model Design and Training Approach

The enhanced model structure for the sarcasm identification challenge is based on the RoBERTa model. The multi-head attention method (Fig. 1) is used to enhance the understanding of the context in the text parts. Avoiding overfit is important things so layer Norm and dropout rate 0.3 are used. Which helps in stability during the training process. The output of the RoBERTa model is classified using a classifier. The weighted loss function [0.7, 1.3] is used to enhance the model to take care of class balance in the dataset. The optimization process is carried out using the AdamW optimizer with a learning rate of 1e-5. For the fine-tuning of the training process, the cosine annealing schedule with warm restarts is used. This schedule gradually decelerates the learning process and restarts it at regular intervals. This helps the model avoid the local optima and converge better. The above-discussed strategies are the most optimized and balanced approach for the

problem of sarcasm detection. The proposed SarcasmBERTa model process is illustrated in Figure. 2. Raw data collection marks the first step in the process. Raw data will then be subjected to the data preprocessing step where it undergoes several procedures such as data cleaning through the process of lemmatization, stemming, replacement of emojis with texts, among others. After going through the data preprocessing stage, the data will then be tokenized using the RoBERTa Fig. 1 model’s tokenizer. This model employs the K-Fold Cross-Validation technique for training, and the tokenized input data will be replicated five times.

Tokenizing the Input Text

Let’s take an input text T , to process it, we first tokenize it using a pre-trained tokenizer which is the RoBERTa tokenizer. The tokenizer converts the text into a sequence of integer token IDs, ensuring a fixed-length L using padding or truncation. This fixed-length token sequence becomes the input to the model. Here, X is a function of T , representing a sequence of token IDs produced by the tokenizer.

$$X = \text{Tokenizer}(T) = \{x_1, x_2, \dots, x_L\}, \quad x_i \in \mathbb{Z} \quad (3.1)$$

where x_i represents the token ID for the i -th token in the sequence.

Generating Contextual Embeddings with RoBERTa

As we get the tokenized input, X it is passed through the pre-trained RoBERTa model. It produces contextual embeddings for each token, incorporating the context of the surrounding tokens. These embeddings form a matrix H representing the entire input sequence.

$$H = \text{RoBERTa}(X, M) = \{h_1, h_2, \dots, h_L\}, \quad h_i \in \mathbb{R}^d \quad (3.2)$$

Here $H \in \mathbb{R}^{L \times d}$ is the hidden state matrix. $M \in \{0, 1\}^L$ is the attention mask indicating valid (non-padded) tokens. The RoBERTa hidden layer has a dimensionality of d .

Refining Representations via Multihead Self-Attention

While the conventional attention mechanism in the RoBERTa transformer architecture is already employed, an additional MHA layer is added on top of

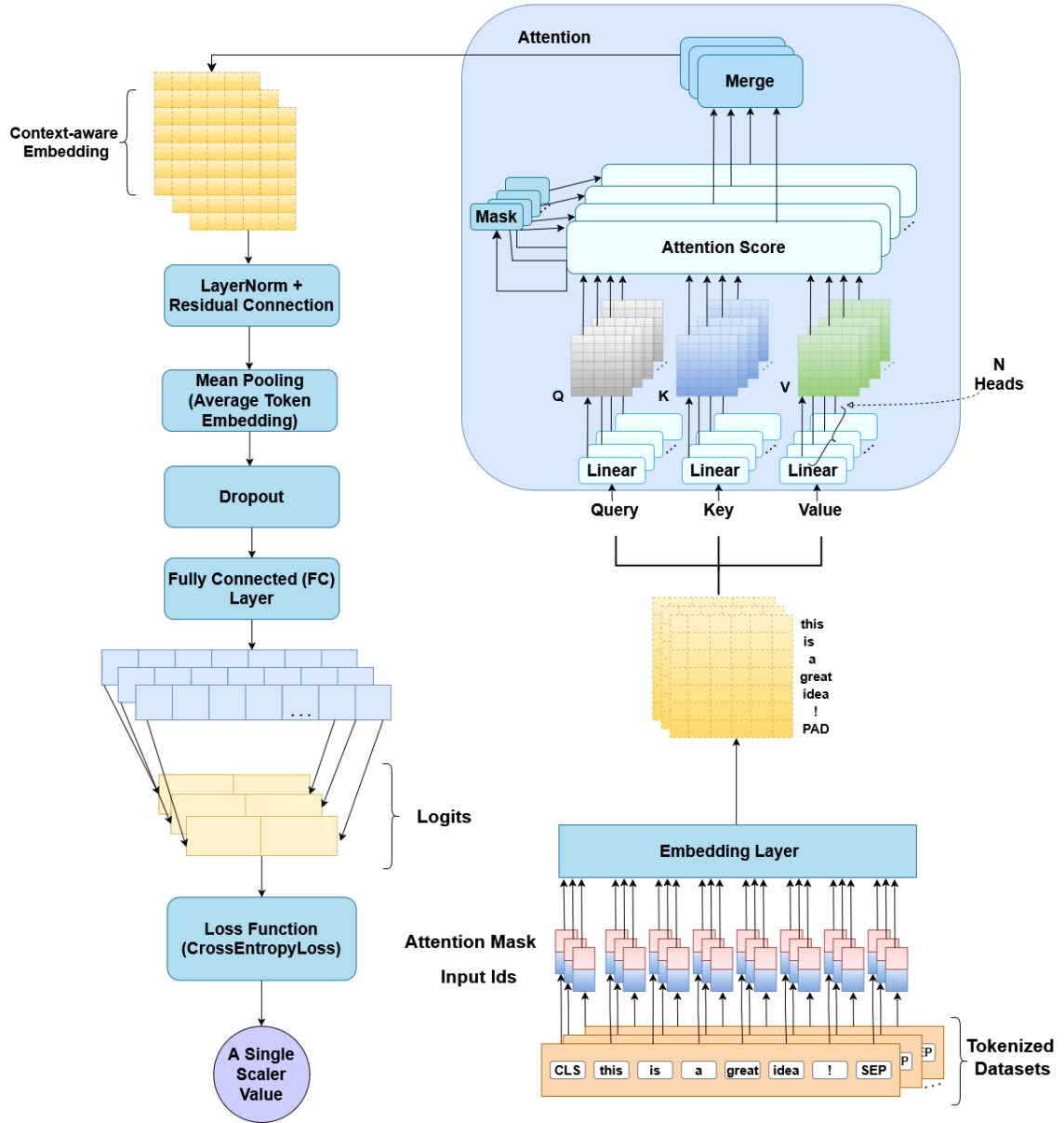


Figure 3.1: Architecture of the proposed sarcasm detection SarcasmBERTa model

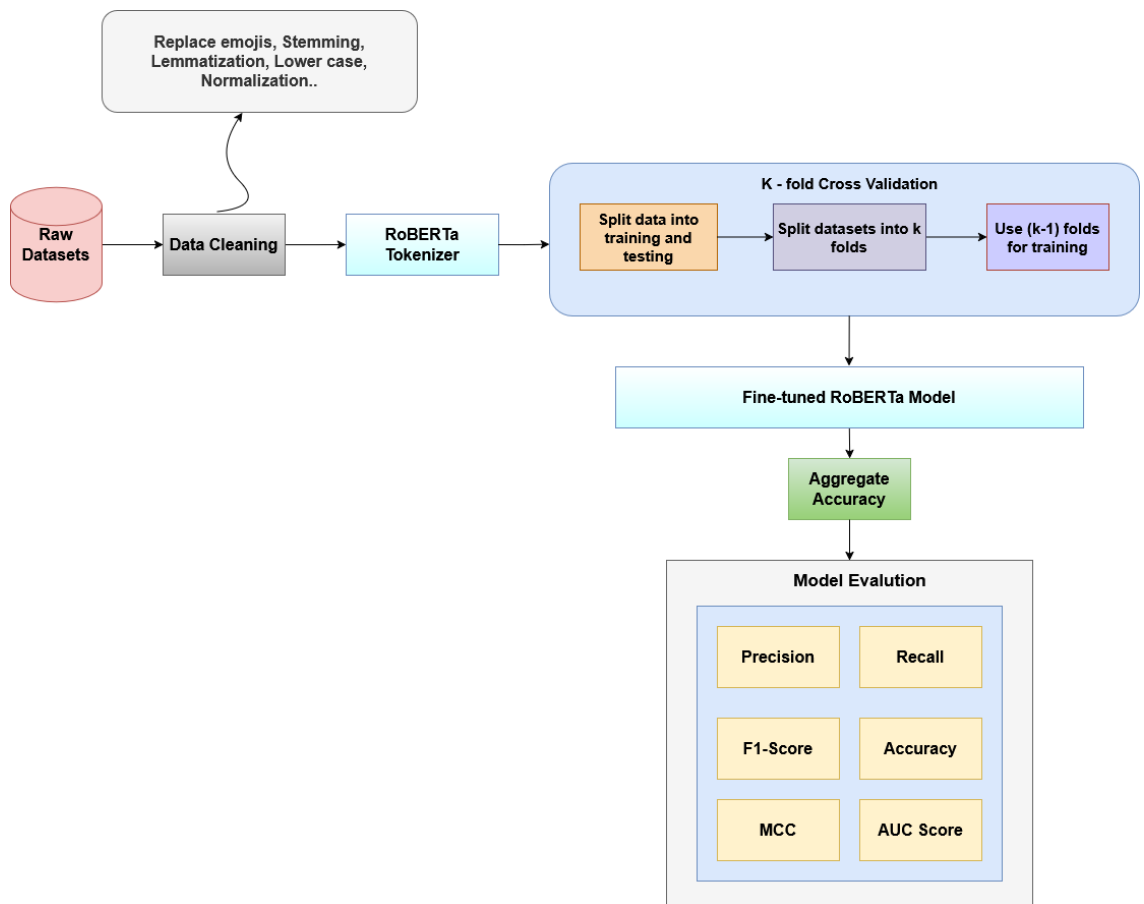


Figure 3.2: Workflow of the Sarcasm Detection Model, i.e., SarcasmBERTa

the context-aware representations to enhance token-level interactions specific to sarcasm contexts. Multihead attention is a good approach for contextual understanding in text. So, to further enhance the contextual representations, a multihead self-attention mechanism is applied.

All attention head outputs are concatenated, which allows the model to project the resulting vector back into the original embedding space.

$$H' = \text{Concat}(\text{Attention}_1, \dots, \text{Attention}_k)W_O \quad (3.3)$$

where $W_O \in \mathbb{R}^{d \times d}$ is the output projection matrix.

To enhance the stability and efficiency of training, we add a residual connection into the model architecture.

$$H'' = \text{LayerNorm}(H' + H) \quad (3.4)$$

Mean Pooled Sequence Representation

Mean pooling is done on the token embeddings in order to obtain a fixed-size representation of the sequence. The mean pooling works by averaging the values of the token embeddings, ensuring the resulting vector captures the sequence’s overall characteristics. Moreover, it minimizes the effect of single tokens, making it possible for the model to pay attention to the whole sequence. This technique proves efficient in obtaining a brief and fixed-size representation of the sequence.

$$P = \frac{1}{L} \sum_{i=1}^L h_i'' \quad (3.5)$$

The resulting vector $P \in \mathbb{R}^d$ serves as the global representation of the input sequence.

Classification Layer for Sarcasm Detection

To avoid overfitting, a dropout layer is used on the pooling vector P . Another dense layer, which is completely connected, processes the vector after dropout. This layer transforms the pooled vector into logits, which are class-specific scores. It enables the model to make predictions on all classes. To increase the robustness of prediction, we use dropout before classifying, and

a fully connected layer provides raw prediction values for both sarcasm and non-sarcasm labels.

$$Z = W_F P + b, \quad Z \in \mathbb{R}^C \quad (3.6)$$

where $W_F \in \mathbb{R}^{d \times C}$ is the weight matrix, $b \in \mathbb{R}^C$ is the bias vector, $C = 2$ is the number of output classes (Sarcastic, Non-Sarcastic). The softmax function is used to forecast the probabilities of the classes:

$$\hat{y} = \text{softmax}(Z) \quad (3.7)$$

Training with Weighted Cross-Entropy Loss

During training, a weighted cross-entropy loss function is used to handle the imbalance of the dataset:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N w[y_i] \log(\hat{y}_{i,y_i}) \quad (3.8)$$

where y_i is the true label of the i -th sample, $\hat{y}_{i,y_i}$ is the predicted probability of the true class, $w[y_i]$ is the class weight for y_i , N is the batch size.

Optimizing with AdamW and Learning Rate Scheduling

Model parameters are optimized using AdamW optimizer, which is a variant of Adam that decouples weight decay from the d=gradient update. In addition to the optimizer, we employ a learning rate scheduler to control the pace of learning over time. Initially, a warmup phase is applied to gradually increase the learning rate, which helps stabilize early training and prevent large, unstable updates. This combination of AdamW and learning rate scheduling improves both training efficiency and final accuracy.

$$\eta_t = \eta_0 \cdot \text{cosine_annealing}(t) \quad (3.9)$$

where η_0 is the initial learning rate and t is the current training step.

Evaluating with K-Fold Cross-Validation

The K-Fold Cross-Validation technique is utilized to ensure reliability and universality because the model is trained using $K - 1$ folds and tested using

the last fold. In addition, among those included in the evaluation metrics are Loss, F1-Score, and Accuracy. The mean of the parameters obtained from each fold represents the final result.

Dataset and Evaluation Metrics

There are a total of five datasets. These datasets are for English, Low Resource Language, Hindi, and Bengali. The descriptions of the datasets are as follows:

English

33,859 English tweets are annotated as 1 for sarcasm and 0 for non-sarcasm. The data is cleaned up by removing columns, hence only retaining the required columns. There are 17,598 non-sarcastic tweets, and 16,261 tweets are sarcastic [10].

Bengali.

Lora et al. [33] create a dataset 'Ben-Sarc' (A Dataset for Sarcasm Detection), 25,636 balanced datasets combination of equal division of sarcastic and non-sarcastic datasets.

Hindi

Katyayan introduced [24] a collection of a total of 16,000 Hindi tweets, where 6,051 tweets are sarcastic and 10,128 tweets are non-sarcastic. Kumar et al. include [28] "Panchayat" – a Hindi web series from HumourHindiNet, which offers 3,000+ humor-rich dialogues. "Kota Factory" – another Hindi web series from HumourHindiNet – consists of 2,000+ sarcastic dialogues.

Evaluation Metrics

In the current section, we provide justification for various metrics used for their performance measures based on our recommended algorithm. The explanations for these metrics are mentioned below:

1. **Accuracy (Acc):** Accuracy reflects the ratio of correctly predicted instances out of the total. In simple terms, it tells you how often the

predictions are right. Mathematically, it is:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (3.10)$$

2. **Precision (Pre):** Precision tells us how many of the things we call positive are actually correct. In simple terms, it shows the accuracy of our positive predictions. Mathematically, we write it as:

$$Precision = \frac{TP}{TP + FP} \quad (3.11)$$

3. **Recall (R):** Recall is a performance metric that calculates the ratio of genuinely positive instances labeled correctly. Recall can be formulated mathematically for a binary classification problem as:

$$Recall = \frac{TP}{TP + FN} \quad (3.12)$$

4. **F1 Score (F1):** Harmonic mean (F1 Score) refers to the compromise between recall and precision. This means that the performance of the model is determined by taking both accuracy and ability to capture all positive examples into consideration. The formula used to determine it is:

$$F1 \text{ Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3.13)$$

5. **Consistency Loss (CL):** Consistency Loss calculates the reliability of predictions made by the model in case of minor changes in the input text. For sarcasm detection models, minor changes in word choice and representations should not affect the output class predictions, provided that the semantics remain unchanged. Thus, consistency loss is the measure of dissimilarity between the probability distributions predicted by the model on the input and its perturbation. Let x be the input text, and let $\tilde{x}$ denote a perturbation of the same text obtained using the method of random dropout through forward propagation. The probability distributions corresponding to the respective inputs are denoted as $p(x)$ and $p(\tilde{x})$. This loss is shown below:

$$CL = ||p(x) - p(\tilde{x})||^2 \quad (3.14)$$

When the consistency loss takes a smaller number, this means that the model provides consistent predictions against small disturbances, which suggests that it is more robust.

6. **MCC:** It ranges from -1 for inverse prediction to $+1$ for perfect prediction in the case of binary classification. A perfect prediction is shown by $+1$, an average random guess is shown by 0 , and an inverse prediction is shown by -1 . Matthews Correlation Coefficient (MCC) is particularly useful in cases of unbalanced classes because it considers the imbalance in the calculation. Its mathematical expression is:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.15)$$

7. **ROC Curve (Receiver Operating Characteristic Curve):** is a graphical representation of the performance of a binary classification model for all classification thresholds. Using the ROC curve, we can now also compare different classification methods. Mathematically it is written as:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \quad (3.16)$$

3.1.2 Mental Health Detection Framework

Data Collection

Experiments have been performed using the Reddit dataset, called the Reddit Suicide Watch and Mental Health (SWMH), which consists of a number of mental health-focused subreddits from the Reddit platform [21]. The dataset primarily consists of posts related to issues like depression, anxiety, and bipolar disorder. These are all mental health based. The dataset contains a total of 54,412 labeled posts, which makes it a complete dataset for mental health classification tasks. The dataset has five tags: Anxiety, bipolar, depression, SuicideWatch, and offmychest. Also all tags having prefix "Self.". It is divided into a data to train, data for validation, and data to test in a ratio of 70:10:20. The labels have been remapped to Anxiety, Bipolar, Depression, Suicide, and None, respectively, for better interpretation of the results. The OffMyChest category has been included as a neutral category, which does

not indicate any mental health disorder. The posts included in this category express emotional content but do not indicate any particular mental health disorder.

Data Processing

The social media texts on mental health issues might contain informal writing, irrelevant information, and contradictions. To guarantee high-quality input, the following preprocessing steps are taken:

- Noise removal: removal of URLs, HTML tags, mentions, hashtags, and other non-text symbols
- Normalization: converting all input characters to lowercase and limiting excessive character repetition (e.g., replacing “soooo” with standard whitespace normalization)
- Text standardization: unicode normalization and stop word filtering, as appropriate
- Label encoding: applying a multi-class encoding scheme on mental health categories, which includes:
 - Depression
 - Suicide Watch
 - Disorder Off (Note: The category names can be modified according to the exact categorization scheme used)

Knowledge Graph Construction and Integration

To enrich contextual representations with structured semantic knowledge, we incorporate a domain-specific knowledge graph (KG) representing relationships among mental health concepts. The knowledge graph consists of entities such as symptoms, emotional expressions, psychological states, and mental health conditions extracted from mental health lexicons and Reddit mental health discussions.

Each entity in the knowledge graph is represented using the TransE embedding method, which learns low-dimensional vector representations by modeling relationships between entities as translation operations in the embedding space. Formally, the embedding model optimizes the relationship:

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t} \quad (3.17)$$

The head and tail entity embeddings are denoted by $\mathbf{h}$ and $\mathbf{t}$, respectively, and $\mathbf{r}$ represents the relation embedding vector.

The resulting knowledge graph embeddings provide semantic representations of mental health concepts that complement the contextual embeddings produced by DistilBERT. These embeddings are then integrated with textual representations during feature fusion by enhancing the capturing ability of model implicit psychological signals. The contextual representation obtained from DistilBERT and the knowledge graph embedding are fused using feature concatenation:

$$\mathbf{h}_{\text{fused}} = [\mathbf{h}_{\text{text}}; \mathbf{h}_{\text{KG}}] \quad (3.18)$$

where $\mathbf{h}_{\text{text}}$ represents contextual embeddings generated by DistilBERT and $\mathbf{h}_{\text{KG}}$ represents knowledge graph embeddings.

Large Language Models

Emotional well-beings prediction problems are evaluated by a number of text-based (language) models to check effectiveness

- BERT: This model was fine-tuned to apply the model to the SWMH dataset.
- RoBERTa: This is an optimized version of the BERT model. It was fine-tuned to enable performance comparisons with the baseline model.
- MentalBERT and MentalRoBERTa: These are domain-specific transformer models that were trained on text data related to mental health. They were used directly.
- MentaLLaMA: This is a language model based on the LLaMA2 model. Evaluation was done in predicting mental health outcomes.
- BART-Large-MNLI: Zero-shot classification was used to predict mental health classes independently.

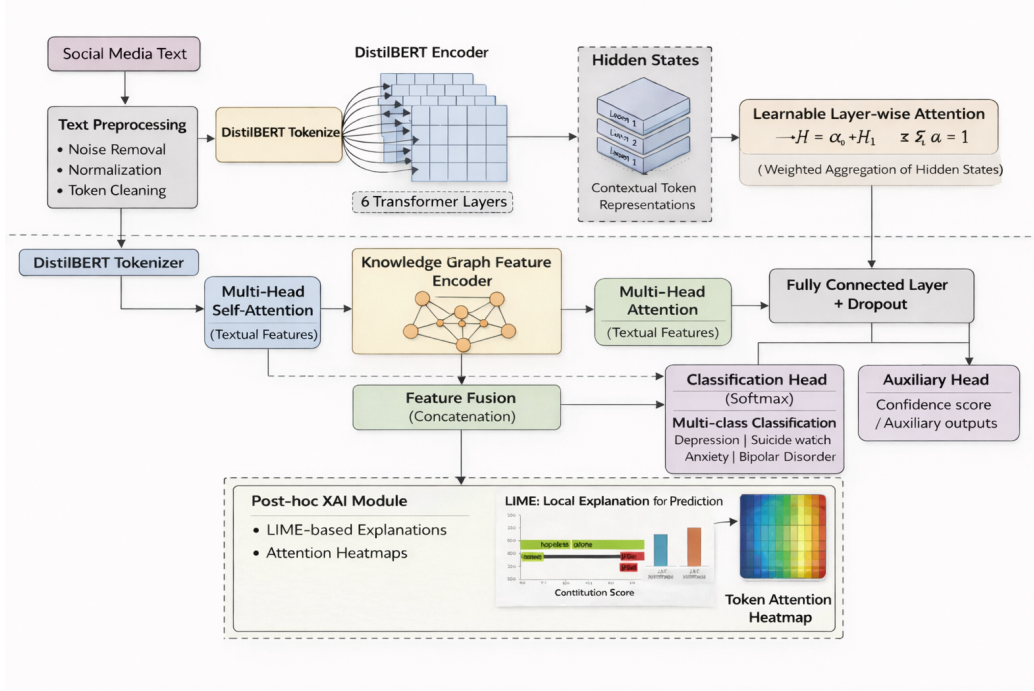


Figure 3.3: Overall architecture of the proposed X-MHD model integrating DistilBERT, Knowledge Graph features, and LIME-based explainable AI for multi-class mental health disorder detection.

Evaluation Metrics

For evaluating the model comprehensively, various metrics were used:

- **Accuracy:** This metric calculates the overall accuracy of predictions made by the model.
- **Classification Report:** This metric provides detailed information regarding precision, recall, and F1 score for different classes.
- **Confusion Matrix:** Calculation of true positives, true negatives, false positives, and false negatives is done. These are needed to create confusion matrix.

Multi-Task Learning Framework

To improve model robustness and generalization, we adopt a multi-task learning framework consisting of a primary classification task and an auxiliary prediction task. The primary task performs multi-class classification of mental health conditions, while the auxiliary task predicts a confidence score indicating whether the input text expresses mental health related distress.

The overall training objective is defined as a weighted combination of the two loss functions:

$$L_{\text{total}} = L_{\text{main}} + \lambda L_{\text{aux}} \quad (3.19)$$

where L_{main} represents the multi-class classification loss and L_{aux} denotes the auxiliary task loss. The parameter λ controls the contribution of the auxiliary task during training.

Hyperparameter Settings

The proposed X-MHD model employs *DistilBERT-base-uncased* as the pre-trained language encoder, consisting of six transformer layers. Maximum sequence length of 256 is used to tokenize input texts. In this model batch size is 32, AdamW optimizer at a learning rate of 1×10^{-5} . To reduce overfitting, a dropout rate of 0.3 was applied to fully connected layers. The multi-head attention module utilizes eight attention heads, while the knowledge graph encoder generates embeddings of dimension 50. A class-weighted cross-entropy loss is adopted to address class imbalance, complemented by an auxiliary loss weighted at 0.3 to improve training stability. The model is trained for 10 epochs with 100 warm-up steps, and performance’s evaluation is done using 5-fold cross-validation, with all hyperparameters kept fixed across folds to ensure reproducibility.

Chapter 4

RESULT & ANALYSIS

4.1 Results for Sarcasm Detection

Here, we explained the experimental comparison of the methodology developed. We performed experiments on the data sets given in Section 4, and performance was measured by metrics as mentioned in Section 5. This enabled us to evaluate how well our proposed method identifies negative stances in multilingual social media posts. In addition, we compared our approach with several baseline algorithms, including TextCNN [14], and Bi-LSTM [17], to gain a more comparative study.

Evaluation of the model was performed against a variety of existing modern-day sarcasm detection models, including those proposed by Zhang et al. [53], Tay et al. [50], Devlin et al. [26], Xiong et al. [53], and Nguyen et al. [38]. Table 1 illustrates the hyperparameters used to enhance the performance of the model. The batch size used in this case was set to 16 for efficiency, and the max length was set to 256 to ensure context. The learning rate was set to 1e-5, and the model was trained for 50 epochs. For efficient training, 100 warm-up steps were used. The dropout rate was set to 0.3 to ensure that overfitting was avoided. For the multi-head attention, the number of heads was set to eight. The model used the RoBERTa model to perform the role of a transformer. The model also used multi-head attention to ensure that it effectively extracted all the information from the text. In the context of effective training process stability, the model utilizes the concept of ‘Layer Normalization.’ For the purpose of efficient regularization, the model utilizes regularization. In the context of the optimization process for the fine-tuning phase, the model utilizes the concept of ‘AdamW.’ A ‘cosine annealing’ learning rate scheduler is utilized to efficiently plan the learning rate. In the context of effectively addressing the ‘class imbalance,’ the model utilizes the concept of ‘class weighting.’

4.1.1 Hyperparameter Study

A specific number of parameters is used, which include batch size, epochs, warmup steps, learning rate, etc. In this section, the hyperparameters of the model are examined. Max length represents the input token's maximum sequence length. When this value is increased, Batch size is needed when max length become high. When the batch size is high, it enables the model to learn better without using up a lot of memory. The number of times the model iterates over the entire dataset is determined by epochs, which helps in preventing overfitting using early stopping or validation monitoring. The gradual increase in the learning rate using the number of steps in warmup is used to prevent sudden changes in the model weights. When a lower learning rate is chosen, Fine-tuning of pre-trained models like BERT or RoBERTa, which is beneficial in retaining the knowledge that is already learned in the model. Dropout is a regularization method where some of the input elements are randomly set to zero during training. This is beneficial in preventing overfitting by not letting the model get too dependent on certain neurons or features. Representational ability is improved in the model through the use of attention heads, enabling it to pay attention to various parts of the input at once. Class imbalance is addressed by assigning weights to each class within the loss function in order to improve classification performance. AdamW is an optimizer variant that stands for Adam with Weight Decay. It is beneficial in dealing with L2 regularization or weight decay, especially in transformer-based models like BERT, RoBERTa, etc. The scheduler is used to control the learning rate during training, which is beneficial in helping the model converge faster with high accuracy.

4.1.2 Performance comparison with Baseline Algorithms and Contemporary Models

This section sought to assess how well the proposed algorithm performed in comparison to a number of baseline algorithms, including TextCNN, Bi-LSTM, and GloVe. To achieve this, the baseline algorithms are refined for optimal performance. The performance of the proposed algorithm was compared to a number of existing algorithms as part of the experimental study, such as those by P. Zhang et al. [53], Y. Tay et al. [50], J. Devlin et al. [26], and D. Q. Nguyen et al. [38]. Table 3 shows the Correct Rate, Exactness, Sensitivity results, and Harmonic Mean respectively for these algorithms and

ours over the entire set of datasets 1. Analysis of these results demonstrates that our algorithm ranks highest in terms of performance across all datasets and metrics. In Dataset 1, for the English datasets we used Matthews’ correlation coefficient (MCC), for each test fold. The model achieved consistently high MCC scores. The highest MCC score was 0.9211, reflecting the model accurately identified sarcastic and non-sarcastic cases. Consistency Loss is calculated according to how varied the probabilities of the original text and its altered form are when making a prediction. Given the importance of the context in understanding sarcasm, a good sarcasm detection model must have consistent predictions even after minor changes made to the original input data. This means that smaller CL values will show more consistent predictions. Furthermore, to compare the model with multimodal and hybrid approaches presented recently, MMFN [39] and KFGC-Net [60] were evaluated.

Table 4.1: Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Dataset 1 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).

Dataset	Method	Acc	Pre	R	F1	CL
English	TextCNN (Chen, 2015)	0.80	0.74	0.76	0.75	0.052
	SIARN (Tay et al., 2018)	0.80	0.75	0.75	0.76	0.049
	SMSD (Xiong et al., 2019)	0.81	0.76	0.75	0.76	0.047
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.82	0.77	0.78	0.77	0.046
	GloVe (Pennington et al., 2014)	0.82	0.81	0.82	0.81	0.045
	BERT (Devlin et al., 2019)	0.84	0.79	0.82	0.80	0.041
	BERTweet (Nguyen et al., 2020)	0.88	0.87	0.88	0.88	0.034
	KnowleNet (Yue et al., 2023)	0.89	0.88	0.84	0.86	0.029
	KFGC-Net (Zhuang et al., 2025)	0.72	0.66	0.70	0.68	0.059
	MMFN (Ou and Li, 2025)	0.78	0.78	0.76	0.77	0.049
	SarcasmBERTa	0.97	0.96	0.96	0.96	0.019

4.1.3 Evaluation on Bengali Datasets

To evaluate our model’s performance in Bangla datasets, SarcasmBERTa was used. The capability and scalability of Sarcasm detection in a unique language and culture are provided. After performing the experiments on

our model, it provided high accuracy and precision rates. The experiments performed on our model showed its robustness. Most importantly, it was capable of extracting Sarcasm cues from English and Bangla texts effectively, overcoming the hurdles of different language structures and culture. These experiments and results validate our approach and highlight the scalability of our approach for multilingual Sarcasm classification. The model performed greatly, achieving remarkable results despite syntactic and semantic differences from English. These results validate our approach and highlight its scalability for multilingual Sarcasm detection. In this paper, Indic-Transformers Bengali BERT [20] is also referred to as a short form of ‘I-T Bengali BERT’. Table 3 showed the comparison with other models and our proposed model.

Table 4.2: Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Dataset 2 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).

Dataset	Method	Acc	Pre	R	F1	CL
Bengali	TextCNN (Chen, 2015)	0.65	0.64	0.65	0.64	0.061
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.66	0.65	0.66	0.65	0.059
	GloVe (Pennington et al., 2014)	0.66	0.66	0.66	0.66	0.058
	BERT (Devlin et al., 2019)	0.68	0.69	0.64	0.66	0.054
	I-T Bengali BERT (Jain et al., 2020)	0.75	0.74	0.77	0.75	0.041
	KFGC-Net (Zhuang et al., 2025)	0.70	0.68	0.69	0.68	0.058
	MMFN (Ou and Li, 2025)	0.73	0.72	0.72	0.72	0.052
	SarcasmBERTa	0.82	0.82	0.82	0.82	0.026

4.1.4 Evaluation on Hindi Datasets

SarcasmBERTa’s ability to detect sarcasm across different languages was examined using the Hindi Tweets Dataset-3 for Sarcasm Detection, which contained over 16,000 tweets. The dataset [24] included 6,051 sarcastic tweets and 10,128 non-sarcastic tweets, making it a rich resource for sarcasm detection in Hindi. We used this dataset to test SarcasmBERTa’s ability to generalize across the linguistic and cultural aspects of Hindi. The efficiency of SarcasmBERTa in recognizing sarcasm in various conditions was evident from its high accuracy and precision rate. The effectiveness of adapting the

model to the unique syntax and semantics of the Hindi language was evidenced through the results, where the effectiveness of using the approach for detecting sarcasm across multiple languages was clearly demonstrated. It was observed that SarcasmBERTa performed excellently on the dataset. This is came to notice that, SarcasmBERTa performed exceptionally well on the data set, and the results were great. The results proved that this approach was sufficiently powerful to detect sarcasm in various complex contexts. In addition, the results proved that the model had the potential to adapt well to the specific syntactic and semantic characteristics of the Hindi language. The results shown in Table 5 proved that the model was generalizable and successful in multilingual sarcasm detection. We used two additional datasets: Dataset 4 to further evaluate the performance of SarcasmBERTa, which was based on the Hindi web series *Panchayat*, and Dataset 5, based on *Kota Factory*. The data sets used were for the purpose of humor detection, with utterances being classified as either ‘Humor’ or ‘Non-Humor.’ The data sets selected were appropriate because they relate to sarcasm and humor, allowing us to assess SarcasmBERTa’s ability to differentiate between them. The robustness of the model in identifying sarcasms was affirmed upon scrutinizing the data sets, which indicated that the model had excellent accuracy and precision in identifying them in both cases. The comparison of the performance of our proposed model, Sarcasm-BERTa, and other models, both old and new, was demonstrated in Table 4. SarcasmBERTa demonstrated its efficiency in handling all datasets, including English, Bengali, and Hindi.

The evaluation parameters to measure the efficacy of our model, as shown in Table 5, were:

Figure 3 shows loss curves, that there is a general trend of decreasing loss throughout the 50 epochs. Within the first 10 epochs, the loss decreased rapidly from around 0.5 to less than 0.2. This shows that the network is learning and adapting very well. At the same time, the validation loss is also decreasing, albeit with some fluctuations. In the second part of the graph, between epochs 11 and 30, the network continues to reduce its loss steadily, now to below 0.1. At the same time, the validation loss is steady between 0.2 and 0.3. This shows that the network is becoming more consistent in its performance. Finally, in the third part of the graph, between epochs 31 and 50, the loss is very low, now below 0.05, and the validation loss is also low at around 0.2. This shows that the network has been able to minimize its errors without the risk of overfitting.

During the first phase, from epochs 1 to 10, the training accuracy rose



Figure 4.1: Representation of Training and Validation Loss over Epochs

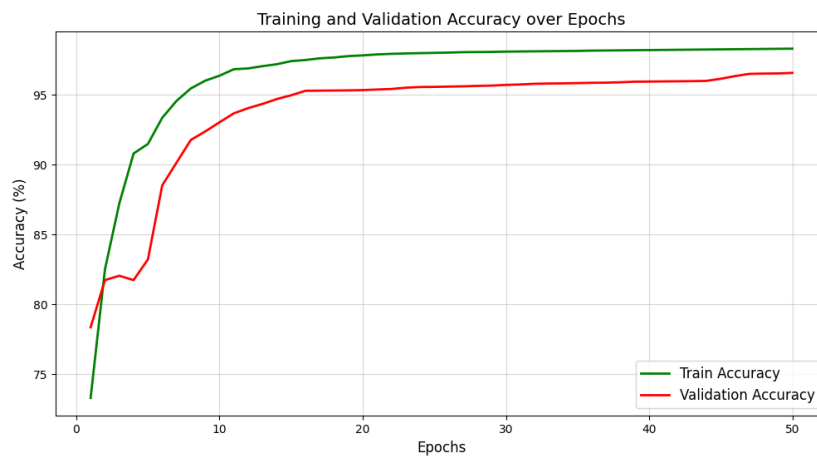


Figure 4.2: Representation of Training and Validation Accuracy over Epochs

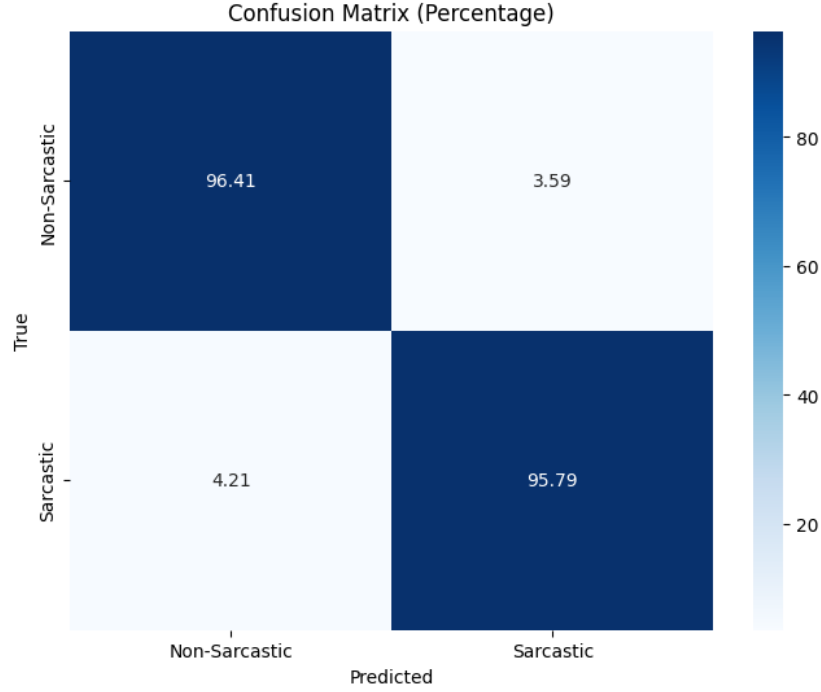


Figure 4.3: Confusion Matrix for Sarcasm Detection Model

from around 73% to over 90%. At the same time, validation accuracy increased from 78% to about 88%, with only minor fluctuations. In the middle phase, from epochs 11 to 30, the training accuracy exceeded 96%, while the validation accuracy grew steadily, crossing 90%. The gap between the two metrics was very small, indicating that the model generalized well. In the final phase, from epochs 31 to 50, the training accuracy reached nearly 98%, and the validation accuracy reached 96.10% at the end of the 50 epochs. The convergence of both metrics confirmed improvement of our model and the absence of overfitting. The confusion matrix (Fig. 5) showed high level of performance of our proposed model. Both the TP and TN values were well above 90%. Specifically, it was observed that the model correctly classified 95.79% of sarcastic instances and 96.41% of non-sarcastic ones. The FP and FN rates were also very low, at 3.48% and 4.84%, respectively. Misclassifications rates between sarcastic and non-sarcastic texts were minimal, reflecting its competence in recognizing nuanced language patterns.

4.1.5 Ablation Study

An ablation research was carried out with the aim of focusing on the effect of the suggested custom multi-head attention mechanism in order to effectively evaluate the contribution of the suggested components. For this purpose, different improved versions with the incorporation of attention processes were compared with the suggested fine-tuning of the RoBERTa method. The aim behind this research is to evaluate the effect of the suggested custom multi-head attention layer on the model’s ability to effectively recognize contextual sarcastic signals in multilingual datasets. In addition, a consistency regularization term was incorporated during training to encourage stable predictions for perturbed inputs. This regularization further improved SarcasmBERTa (RoBERTa + Custom Multi-Head Attention) model. Table 4.4 shows the ablation research results.

4.2 Results for Mental Health Detection

4.2.1 Ablation Study

An ablation study is done. We measured the performance of our model using accuracy, precision, recall, and F1-score. as shown in Table I, fine-tuned BERT-base and RoBERTa achieved accuracies of 71% and 72%, respectively, while domain-adapted models such as MentalBERT and Men- The RoBERTa model achieved a performance of up to 70%. In contrast, large and zero-shot models, such as BART-large MNLI and MentalLLaMA, comparatively lower accuracy, emphasizing their relative ineffectiveness in this task without task-specific adaptation. The proposed X-MHD model, developed using a DistilBERT-based contextual encoder enriched with knowledge graph representations, achieved the highest test accuracy of 88% while maintaining a parameter- efficient architecture. As shown in Fig. 4.4, the LIME explanation highlights emotionally significant tokens such as “suicidal” and “wont”. While less relevant terms prevent negative influence, exemplifying the interpretability of the proposed model. As shown in Fig. 4.5, the model shows excellent class-wise discrimination, with minor confusion among semantically overlapping categories.

Although other explainability methods such as SHAP and Integrated Gradients have been explored in prior work, LIME was selected in our framework

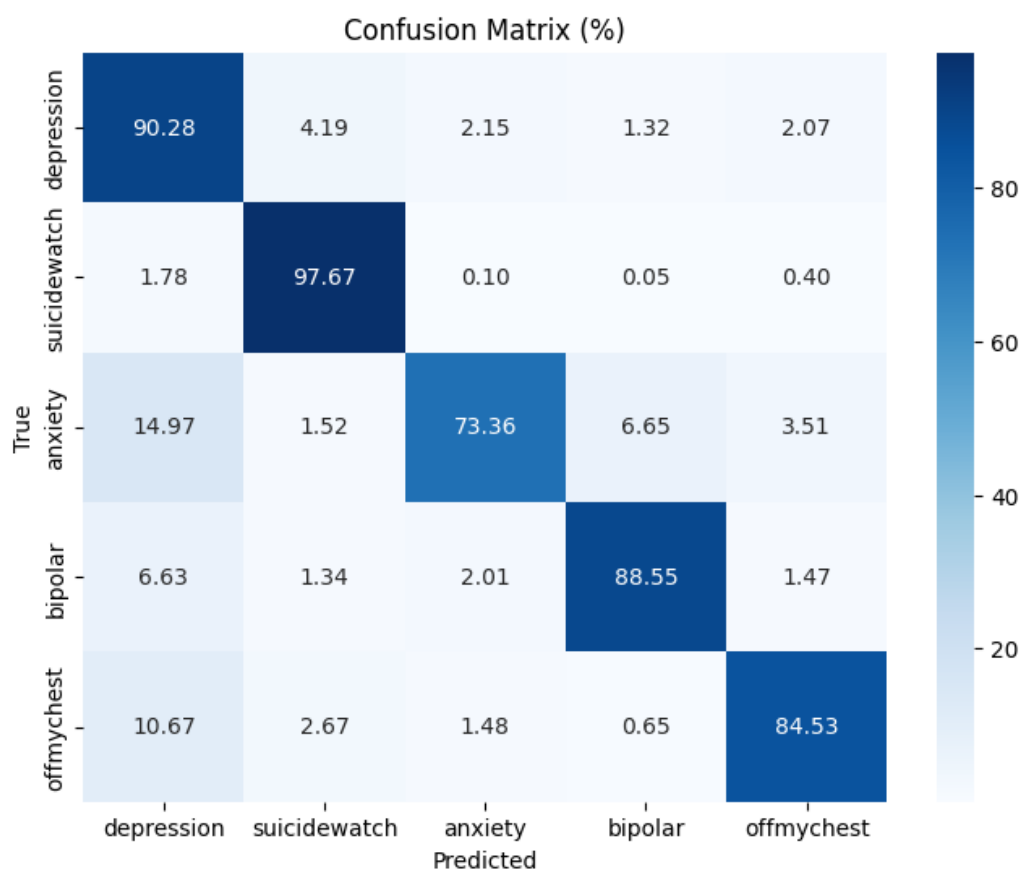


Figure 4.4: F1-score comparison across mental health conditions.

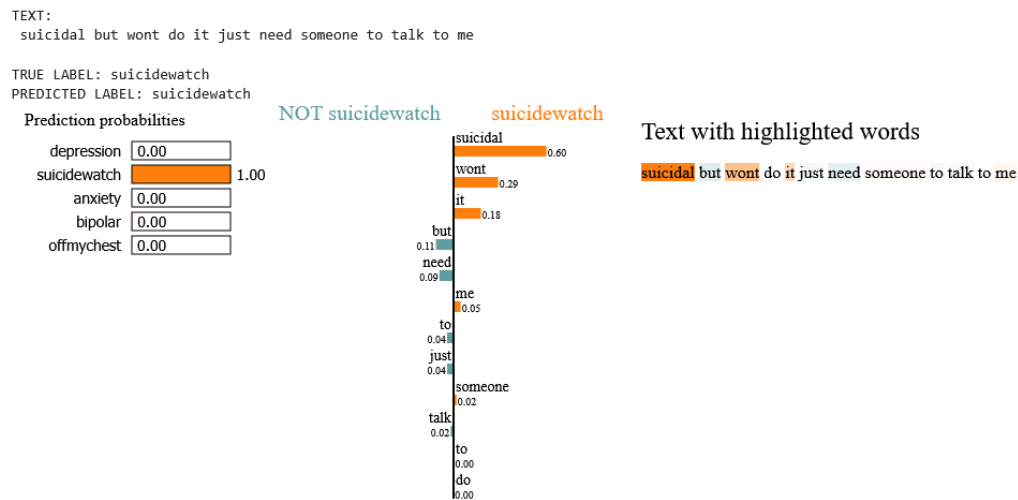


Figure 4.5: LIME-Based Local Explanations for Token-Level Contributions to Predicted Mental Health Category.

as it provides clear local explanations for transformer-based models.

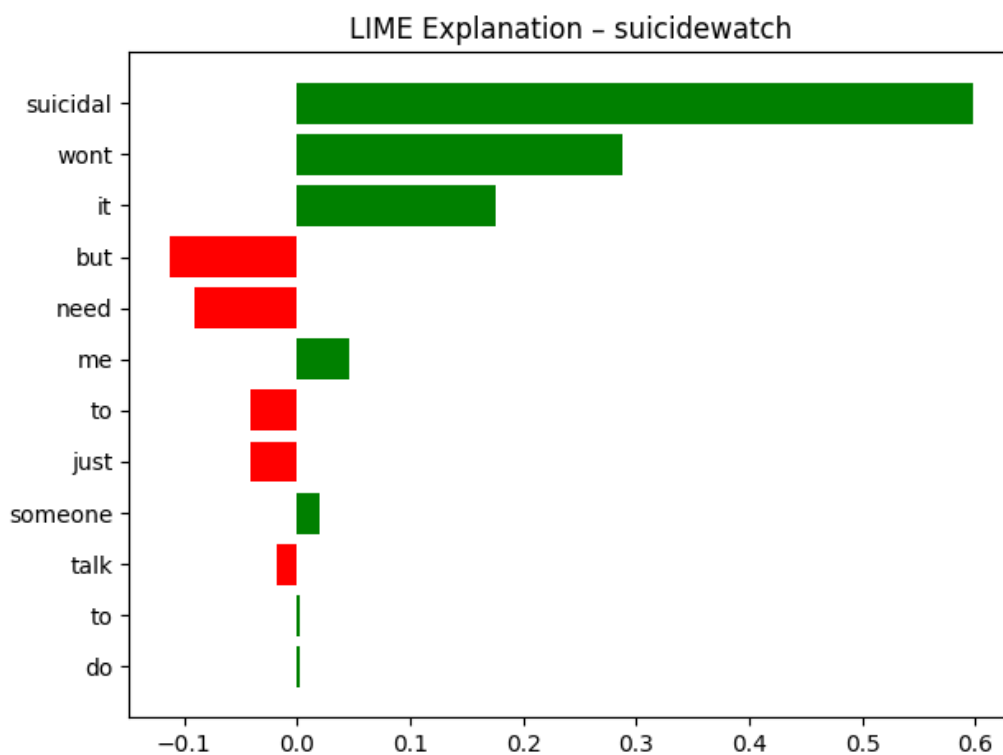


Figure 4.6: LIME-Based Local Explanation for Suicide Watch Prediction, with Token-Level Contributions.

Table 4.3: Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Datasets 3, 4, and 5 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).

Dataset	Method	Acc	Pre	R	F1	CL
Hindi	TextCNN (Chen, 2015)	0.79	0.73	0.75	0.74	0.053
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.80	0.77	0.78	0.77	0.051
	GloVe (Pennington et al., 2014)	0.81	0.78	0.79	0.78	0.048
	BERT (Devlin et al., 2019)	0.82	0.80	0.81	0.80	0.044
	HumourHindiNet (Kumar et al., 2024)	0.84	0.83	0.84	0.83	0.039
	KFGC-Net (Zhuang et al., 2025)	0.78	0.76	0.77	0.76	0.050
	MMFN (Ou and Li, 2025)	0.81	0.79	0.80	0.79	0.044
	SarcasmBERTa	0.96	0.96	0.96	0.96	0.021
Kota Factory	TextCNN (Chen, 2015)	0.78	0.74	0.76	0.75	0.056
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.82	0.77	0.78	0.77	0.052
	BERT (Devlin et al., 2019)	0.89	0.93	0.86	0.89	0.038
	GloVe (Pennington et al., 2014)	0.90	0.95	0.83	0.89	0.036
	HumourHindiNet (Kumar et al., 2024)	0.91	0.96	0.87	0.91	0.033
	KFGC-Net (Zhuang et al., 2025)	0.82	0.79	0.80	0.79	0.047
	MMFN (Ou and Li, 2025)	0.85	0.82	0.83	0.82	0.041
	SarcasmBERTa	0.96	0.92	0.91	0.92	0.022
Panchayat	TextCNN (Chen, 2015)	0.77	0.75	0.77	0.76	0.058
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.80	0.78	0.79	0.78	0.053
	GloVe (Pennington et al., 2014)	0.85	0.83	0.86	0.85	0.046
	BERT (Devlin et al., 2019)	0.86	0.84	0.87	0.86	0.041
	HumourHindiNet (Kumar et al., 2024)	0.87	0.90	0.84	0.87	0.038
	KFGC-Net (Zhuang et al., 2025)	0.84	0.81	0.82	0.81	0.046
	MMFN (Ou and Li, 2025)	0.87	0.84	0.85	0.84	0.039
	SarcasmBERTa	0.94	0.89	0.91	0.90	0.024

Table 4.4: An ablation study to evaluate the impact of the attention mechanism used in the SarcasmBERTa model.

Model Variant	Correct Rate	Harmonic Mean
RoBERTa (Baseline Fine-tuning)	0.92	0.92
RoBERTa + Attention Layer	0.94	0.94
SarcasmBERTa	0.96	0.96

Table 4.5: Ablation Study of X-MHD

Model Variant	Accuracy
DistilBERT only	78%
DistilBERT + Knowledge Graph	83%
DistilBERT + XAI	80%
DistilBERT + KG + XAI	85%
Full X-MHD (Proposed)	88%

Table 4.6: Summary of Experimental Results on the Test Dataset

Model	Parameters	Accuracy
MentalBERT-base	110 M	69%
MentalRoBERTa	110 M	70%
MentalLLaMA-7B	7 B	59%
<i>Fine-Tuning</i>		
BERT-base	110 M	71%
RoBERTa-base	110 M	72%
<i>Zero-Shot Prompting</i>		
BART-large-mnli	400 M	49%
<i>Proposed Model</i>		
X-MHD	60 M	88%

Table 4.7: Per-Class Performance of X-MHD

Class	Precision	Recall	F1-Score
Depression	0.89	0.87	0.88
Anxiety	0.85	0.83	0.84
Bipolar	0.86	0.85	0.85
SuicideWatch	0.90	0.91	0.90
OffMyChest	0.82	0.84	0.83

Chapter 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

5.1 Conclusion

The research addressed important challenges associated with contextual ambiguity, multilingual expressions, emotional understanding, and interpretability in social media analysis. For sarcasm detection, the proposed Sarcasm-BERTa framework integrated RoBERTa with Multi-Head Attention mechanisms to capture contextual and semantic relationships in multilingual social media conversations. The model demonstrated superior performance on English, Hindi, and Bengali datasets compared to several traditional and transformer-based baseline approaches. The experimental results showed that the proposed framework effectively identified implicit sarcastic expressions and improved multilingual sarcasm classification accuracy. For mental health detection, the proposed X-MHD framework combined DistilBERT, knowledge graph embeddings, and Explainable Artificial Intelligence techniques for multi-class mental health classification. The framework successfully identified mental health conditions such as depression, anxiety, bipolar disorder, and suicide-related behaviour from textual data collected from online communities. The integration of contextual representation learning and semantic knowledge significantly improved prediction performance and robustness. Furthermore, Explainable Artificial Intelligence techniques including LIME and attention visualization enhanced the interpretability and transparency of the proposed systems by highlighting important words, contextual cues, and semantic relationships responsible for model predictions. These approaches improved the trustworthiness and practical applicability of the proposed frameworks in healthcare-related applications. The proposed frameworks demonstrated the importance of transformer-based contextual learning, attention mechanisms, semantic feature integration, and explainability for improving social media text understanding.

5.2 Future Scope and Social Impact

Future work may focus on extending the proposed frameworks toward multi-modal learning approaches using textual, audio, visual, and facial expression data for improved emotional and psychological analysis. Additional improvements may include integrating lightweight large language models, real-time mental health monitoring systems, adaptive knowledge graph learning, and low-resource multilingual learning for resource-efficient deployment. The incorporation of reinforcement learning and conversational AI systems may further improve user-aware mental health assistance and personalized recommendation systems.

The social impact of this research is significant because social media platforms contain large amounts of emotional and psychological expressions shared by users worldwide. Early identification of mental health risks and emotional distress through automated AI systems can support healthcare professionals, researchers, educational institutions, and online communities in providing timely intervention and awareness. Similarly, accurate sarcasm understanding can improve sentiment analysis, social media moderation, human-computer interaction, and context-aware conversational systems. Therefore, the proposed research has the potential to contribute toward socially responsible, interpretable, and human-centric AI systems for digital mental health and social media understanding.

Bibliography

- [1] AHMAD, I. A., GATLA, P., AND MUNDOTIYA, R. K. Sarcasm identification and classification in hindi newspaper headlines. *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, 4 (2025), 1–21.
- [2] AHUJA, R., AND SHARMA, S. C. Transformer-based word embedding with cnn model to detect sarcasm and irony. *Arabian Journal for Science and Engineering* 47, 8 (2022), 9379–9392.
- [3] ANAN, R., APON, T. S., HOSSAIN, Z. T., MODHU, E. A., MONDAL, S., AND ALAM, M. G. R. Interpretable bangla sarcasm detection using bert and explainable ai. In *2023 IEEE 13th annual computing and communication workshop and conference (CCWC)* (2023), IEEE, pp. 1272–1278.
- [4] BADE, G. Y., OROPEZA, J. L., AND KOLESNIKOVA, O. Pragmatic generalization in llms: Insights from fine-tuning and evaluating on multilingual sarcasm. In *Mexican International Conference on Artificial Intelligence* (2025), Springer, pp. 218–230.
- [5] BAO, Y., ZHAO, X., ZHANG, P., QI, Y., AND LI, H. Hian: A hybrid interactive attention network for multimodal sarcasm detection. *Pattern Recognition* 164 (2025), 111535.
- [6] BEDI, M., KUMAR, S., AKHTAR, M. S., AND CHAKRABORTY, T. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1363–1375.
- [7] BIKKU, T., JARUGULA, J., KONGALA, L., TUMMALA, N. D., AND DONTIBOINA, N. V. Exploring the effectiveness of bert for sentiment analysis on large-scale social media data. In *2023 3rd International Conference on Intelligent Technologies (CONIT)* (2023), IEEE, pp. 1–4.
- [8] BOUSMAHA, K. Z., HAMADOUCHE, K., DJOUABI, H., AND HADRICHBELGUITH, L. Automatic algerian sarcasm detection from texts and

images. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 7 (2024), 1–25.

- [9] CAI, Y., CAI, H., AND WAN, X. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), A. Korhonen, D. Traum, and L. Màrquez, Eds., Association for Computational Linguistics, pp. 2506–2515.
- [10] CAI, Y., CAI, H., AND WAN, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (2019), pp. 2506–2515.
- [11] CASTRO, S., HAZARIKA, D., PÉREZ-ROSAS, V., ZIMMERMANN, R., MIHALCEA, R., AND PORIA, S. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).
- [12] CHAUHAN, D. S., SINGH, G. V., ARORA, A., EKBAL, A., AND BHATTACHARYYA, P. An emoji-aware multitask framework for multi-modal sarcasm detection. *Knowledge-Based Systems* 257 (2022), 109924.
- [13] CHEN, J. Q., QI, K., ZHANG, A., SHALAGINOV, M., AND ZENG, T. H. Covid-19 impact on mental health analysis based on reddit comments. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2022), IEEE, pp. 2253–2258.
- [14] CHEN, Y. Convolutional neural network for sentence classification. Master’s thesis, University of Waterloo, 2015.
- [15] DWIVEDI, C., AND RAO, J. Multilingual sentiment analysis for detecting mental health problems using a hybrid algorithm combining rnn and bi-lstm. In *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (2023), IEEE, pp. 527–534.
- [16] D’CRUZ, L., DUBEY, V., AND THAKUR, P. Depression prediction from combined reddit and twitter data using machine learning. In *2023 2nd International Conference for Innovation in Technology (INOCON)* (2023), IEEE, pp. 1–5.

- [17] GRAVES, A., AND SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610. IJCNN 2005.
- [18] GREGORY, H., LI, S., MOHAMMADI, P., TARN, N., DRAELOS, R., AND RUDIN, C. A transformer approach to contextual sarcasm detection in twitter. In *Proceedings of the second workshop on figurative language processing* (2020), pp. 270–275.
- [19] HAZARIKA, D., PORIA, S., GORANTLA, S., CAMBRIA, E., ZIMMERMANN, R., AND MIHALCEA, R. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413* (2018).
- [20] JAIN, K., DESHPANDE, A., SHRIDHAR, K., LAUMANN, F., AND DASH, A. Indic-transformers: An analysis of transformer language models for indian languages. *arXiv preprint arXiv:2011.02323* (2020).
- [21] JI, S., LI, X., HUANG, Z., AND CAMBRIA, E. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications* 34, 13 (2021), 10695–10707.
- [22] JI, S., ZHANG, T., ANSARI, L., FU, J., TIWARI, P., AND CAMBRIA, E. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference* (2022), pp. 7184–7190.
- [23] JOSHY, A., AND SUNDAR, S. Analyzing the performance of sentiment analysis using bert, distilbert, and roberta. In *2022 IEEE international power and renewable energy conference (IPRECON)* (2022), IEEE, pp. 1–6.
- [24] KATYAYAN, P. Tweets dataset for sarcasm detection in hindi. <https://github.com/>, 2020. GitHub repository.
- [25] KAUSHIK, B., SHARMA, A., CHADHA, A., AND SHARMA, R. Machine learning model for sentiment analysis on mental health issues. In *2023 15th International Conference on Computer and Automation Engineering (ICCAE)* (2023), IEEE, pp. 21–25.

- [26] KENTON, J. D. M.-W. C., AND TOUTANOVA, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (2019), vol. 1, Minneapolis, Minnesota.
- [27] KHODAK, M., SAUNSHI, N., AND VODRAHALLI, K. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579* (2017).
- [28] KUMAR, A., MALLIK, A., AND KUMAR, S. Humourhindinet: Humour detection in hindi web series using word embedding and convolutional neural network. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 7 (2024), 1–21.
- [29] KUMAR, S., CHAUHAN, A. R., AKHIL, KUMAR, A., AND YANG, G. Resp-boostnet: Mental stress detection from biomarkers measurable by smartwatches using boosting neural network technique. *IEEE Access* 12 (2024), 149861–149874.
- [30] LAN, Z., CHEN, M., GOODMAN, S., GIMPEL, K., SHARMA, P., AND SORICUT, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
- [31] LIANG, B., GUI, L., HE, Y., CAMBRIA, E., AND XU, R. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing* (2024).
- [32] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [33] LORA, S. K., SHAHARIAR, G. M., NAZMIN, T., RAHMAN, N. N., RAHMAN, R., BHUIYAN, M., AND SHAH, F. M. Ben-sarc: A self-annotated corpus for sarcasm detection from bengali social media comments and its baseline evaluation. *Natural Language Processing* 31, 2 (2025), 674–699.
- [34] MADHURI, C. H. R., SANKAR KRISHNA BANDARU, J., SRINU, M., AND ANAND VARDHAN, G. M. Ai-powered mental health screening and support for homeless children. In *2025 AI-Driven Smart Healthcare for Society 5.0* (2025), pp. 115–120.

- [35] MAJUMDER, N., PORIA, S., PENG, H., CHHAYA, N., CAMBRIA, E., AND GELBUKH, A. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems* 34, 3 (2019), 38–43.
- [36] MARRAPU, H. K., MARAM, B., AND REDDI, P. New analytic framework of public mental health prediction using data science. In *2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)* (2022), pp. 1–6.
- [37] MENG, J., ZHU, Y., SUN, S., AND ZHAO, D. Sarcasm detection based on bert and attention mechanism. *Multimedia Tools and Applications* 83, 10 (2024), 29159–29178.
- [38] NGUYEN, D. Q., VU, T., AND NGUYEN, A. T. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200* (2020).
- [39] OU, L., AND LI, Z. Multi-modal sarcasm detection on social media via multi-granularity information fusion. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 3 (2025), 1–23.
- [40] PANDEY, R., AND SINGH, J. P. Bert-lstm model for sarcasm detection in code-mixed social media post. *Journal of Intelligent Information Systems* 60, 1 (2023), 235–254.
- [41] POKHRIYAL, H., AND JAIN, G. Supposititious sarcasm detection and sentiment analysis coping hindi language in social networks harnessing zipf-mandelbrot probabilistic optimisation and perplexity entropy learning. *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, 2 (2025), 1–28.
- [42] PRAMANICK, S., ROY, A., AND PATEL, V. M. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 3930–3940.
- [43] QIN, L., HUANG, S., CHEN, Q., CAI, C., ZHANG, Y., LIANG, B., CHE, W., AND XU, R. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. *arXiv preprint arXiv:2307.07135* (2023).

- [44] REFAI, D., AL-SHAIBANI, M. S., AND AHMAD, I. Is this the best prompt? scoring prompts for arabic nlp across llms. *IEEE Access* 13 (2025), 171468–171492.
- [45] RODRÍGUEZ, A., CHEN, Y.-L., AND ARGUETA, C. Sincere: A hybrid framework with graph-based compact textual models using emotion classification and sentiment analysis for twitter sarcasm detection. *IEEE Transactions on Computational Social Systems* 11, 5 (2024), 5593–5606.
- [46] ROSID, M. A., SIAHAAN, D., AND SAIKHU, A. Sarcasm detection in indonesian-english code-mixed text using multihead attention-based convolutional and bi-directional gru. *IEEE Access* (2024).
- [47] R.S., P., DEEP, H., AND YADAV, D. Mental health support using gen-ai shot prompting technique and vector embeddings. In *2025 International Conference on Data Science and Business Systems (ICDSBS)* (2025), pp. 1–5.
- [48] SARODE, A. R., RANI, M. V., DAS, B., S, U., AND PUNIA, S. An ai-based framework for interpretable mental health literacy segmentation and decision support. *IEEE Journal of Biomedical and Health Informatics* 29, 12 (2025), 8799–8806.
- [49] SHARMA, S., GHANTASALA, G. S. P., GUPTA, V., SARANGI, P. K., R, R. S., AND SUNGHEETHA, A. Analyzing employees’ mental health using machine learning algorithms. In *2025 12th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (2025), pp. 1–5.
- [50] TAY, Y., TUAN, L. A., HUI, S. C., AND SU, J. Reasoning with sarcasm by reading in-between. *arXiv preprint arXiv:1805.02856* (2018).
- [51] VAJRE, V., NAYLOR, M., KAMATH, U., AND SHEHU, A. Psychbert: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2021), IEEE, pp. 1077–1082.
- [52] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

- [53] XIONG, T., ZHANG, P., ZHU, H., AND YANG, Y. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference* (2019), pp. 2115–2124.
- [54] YANG, K., ZHANG, T., KUANG, Z., XIE, Q., HUANG, J., AND ANANIADOU, S. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024* (2024), pp. 4489–4500.
- [55] YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R. R., AND LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
- [56] YUE, T., MAO, R., WANG, H., HU, Z., AND CAMBRIA, E. Knowl-enet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion* 100 (2023), 101921.
- [57] ZHANG, X., CHEN, Y., AND LI, G. Multi-modal sarcasm detection based on contrastive attention mechanism. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I* 10 (2021), Springer, pp. 822–833.
- [58] ZHANG, Y., YU, Y., ZHAO, D., LI, Z., WANG, B., HOU, Y., TIWARI, P., AND QIN, J. Learning multitask commonness and uniqueness for multimodal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence* 5, 3 (2024), 1349–1361.
- [59] ZHANG, Y., ZOU, C., LIAN, Z., TIWARI, P., AND QIN, J. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing* 16, 4 (2025), 2560–2578.
- [60] ZHUANG, X., ZHOU, F., AND LI, Z. Multi-modal sarcasm detection via knowledge-aware focused graph convolutional networks. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 5 (2025), 1–22.

Explainable Multi-Class Mental Health Classification from Social Media Using DistilBERT and Knowledge Graphs

Suman Saha

*Department Of Computer Science and Engineering
Delhi Technological University
New Delhi, 110042, India
sumansaha0178@gmail.com*

Sanjay Kumar

*Department Of Computer Science and Engineering
Delhi Technological University
New Delhi, 110042, India
sanjay.kumar@dtu.ac.in*

Abstract—Detection of mental health disorders in social media text has received a great interest of research attention in the past few years due to the increasing usage of social media with sharing personal information and engaging in peer discussions on mental health. While existing deep learning techniques showed impressive predictive results in recent times, which also operate in a black-box paradigm. This paper presents X-MHD, an efficient and transparent model in multi-class mental health disorder detection by integrating DistilBERT-based text encoder, knowledge graph representations, and post-hoc explainable artificial intelligence techniques. The model is capable of performing multi-class classification. This paper focuses on five different mental health disorders—depression, suicide risk, anxiety, bipolar disorder, and off-my-chest expressions. To ensure transparency in mental health disorder detection, the X-MHD model uses a hybrid approach to explainability by incorporating layer-wise attention analysis, token-level attention visualization, and local explanations based on the LIME technique. Knowledge graph representations have also been incorporated to ensure enhanced contextual reasoning.

Index Terms—Mental Health Detection, DistilBERT, Knowledge Graphs, Explainable AI, Social Media Analysis

I. INTRODUCTION

Mental health is a crucial aspect of human well-being that influences how independent perceive, respond to, and cope with life events. According to global statistics, a large number of people will suffer from emotional well-beings [1]. A combination of biological, psychological, and social issues that can lead to poor mental health, making people more vulnerable to depression, anxiety, and even suicidal thoughts. According to WHO statistics, there are 800,000 suicides annually [2]. Social media sites such as Reddit have mental health communities, or subreddits, where people freely discuss sensitive issues. These conversations lead to a massive amount of text that is replete with emotional expressions, linguistic patterns, and contextual data about the mental health of the users. On the basis of the result internet communities has emerged as an important source for emotional well-being

analysis on on a large scale. It has already been shown in previous studies that social media data can be used to identify mental health problems and infer psychological states. To distill insights from this data, researchers implement Natural Language Processing (NLP) techniques to extract semantic, emotional, and contextual information from the text. The task usually involves text classification, where the inputs are assigned labels or conditions related to mental health. Traditional techniques, such as Logistic Regression, SVM, as well as RNN and LSTM, have been seen to perform promisingly in this space. Despite the promising performance of existing techniques, it has been observed that most of the conventional techniques lack contextual understanding, integration of external semantic knowledge, and transparency in decision-making. These drawbacks in conventional techniques have been seen to limit their application in sensitive mental health and healthcare-related environments. To address the drawbacks in existing techniques, a new hybrid approach, X-MHD, has been proposed in this study as a tool for explaining mental health disorders in computational social systems. The proposed X-MHD model uses the DistilBERT model, which is a text-based (language) model. Hidden layers use a multi-head attention mechanism. To move beyond text-based analysis, the use of external knowledge graph embeddings is leveraged, thus allowing the incorporation of relevant structured semantic and relational knowledge pertaining to mental health. The knowledge-enhanced representation allows the model to derive latent mental health indicators that may not be explicitly described in the text. To increase the effectiveness and robustness of the model, a multi-task learning strategy with auxiliary prediction objectives is utilized, which is particularly beneficial in noisy and imbalanced social media text. Unlike most traditional black-box deep learning models, the current strategy focuses on the importance of interpretability by incorporating explainable artificial intelligence (XAI) components. Attention heatmaps and LIME local explanations provide fine-grained interpretability at token and feature levels,

thus allowing transparent analysis of the text and semantic factors affecting prediction of model. Significance of this is highlighted in building trust in the model and deploying it in real-world scenarios. The following is a brief overview of the key contributions:

- 1) We propose X-MHD, a hybrid approach that combines the transformer approach with knowledge guidance of checking mental health issues from social media text in an explainable manner.
- 2) We enhance the DistilBERT encoder by using multi-head attention on the hidden layers, which helps in the improvement of contextual and psychological representation learning.
- 3) We incorporate knowledge graph embeddings to add external semantic and relational knowledge, which enhances the discriminative ability of the model.
- 4) We use a multi-task learning strategy with auxiliary prediction tasks to improve robustness and generalization performance on multiple mental health tasks.
- 5) We incorporate techniques from explainable AI, such as attention heatmaps and LIME explanations, to make our predictions transparent and interpretable.
- 6) We conduct extensive experiments with cross-validation on standard datasets (mental health) which illustrate the efficacy of the claimed approach contrast to the current methods.

Some of the remaining sections are discussed below: Related work in mental health detection and NLP is discussed in Section 2. The proposed methodology and experimental design, Section 3. The results and discussion, Section 4, the ethical consideration, Section 5. In Section 6 limitation is shown and lastly conclusion with insights of future work in Section 7.

II. RELATED WORK

Application of machine learning (ML) techniques for analyzing mental health has been widely investigated in previous studies. Early research work mainly focused on traditional techniques of ML models to detect emotional and psychological patterns in textual data. Kaushik et al. [3] studied classical algorithms, including SVM, also Logistic Regression, and Multinomial Naïve Bayes, for sentiment analysis on mental health datasets. Focusing on early detection and intervention, D'Cruz et al. [4] developed ML classifiers based on some contexts gathered from Twitter and Reddit to predict depressive tendencies based on user-generated posts. Similarly, Chen et al. [5] introduced the psychological effects of the COVID-19 pandemic through textual analysis of comments on the website Reddit that were sourced from mental-health related datasets, such as anxiety, depression, SuicideWatch, extending this line of research to multilingual environments, Dwivedi and Rao [6] introduced a hybrid architecture that combines Recurrent Neural Networks (RNNs) with Bidirectional LSTM (Bi-LSTM) models for identifying mental health-related signals in multiple languages. Sharma et al. [7] proposed the ML methods for analyzing mental fitness issues in workplace environments. A significant breakthrough in natural language processing

began with the advent of the architecture based on transformer proposed by Vaswani et al [8]. This paradigm shift marked the beginning of influential pretrained text models, become central to mental health and sentiment analysis since. Bikku et al. [9] assessed the performance of BERT on large-scale sentiment analysis on social media data. In a comparative study, Joshy et al. [10] assessed in mental health domain using BERT, DistilBERT, RoBERTa. Ji et al. [11] introduced two model MentalBERT and MentalRoBERTa, which were pretrained on Reddit data from mental health communities. In a related effort, Vajre et al. [12] proposed PsychBERT, a model pretrained on a combination of biomedical literature and mental health-focused social media data. Madhuri et al. [13] introduced a mental health screening approach for homeless children by analyzing responses to the Pediatric Symptom Checklist (PSC) by using a SVM model. Kumar et al. [14] proposed a smartwatch-based deep learning framework for mental stress detection that addresses the limitation of missing respiration rate in wearable sensors. Sarode et al. proposed [15] a decision support model to assess Mental Health Literacy (MHL) by segmenting individuals in academic communities using modifiable predictors. Addressing the difficulties related to data scarcity also the lack of open-source mental health-focused LLMs. MentaLLaMA, an open-source LLM family built upon the LLaMA2 architecture is proposed by Yang et al. [16]. Yadav et al. [17] introduced an AI-powered mental wellness chatbot that leveraged generative AI and embedding-based relevance detection to deliver context-aware and empathetic responses. Marrapu et al. [18] proposed a data science-based model for predicting public mental health conditions using sensor data and healthcare datasets.

Existing studies rarely combine contextual language modeling, structured knowledge integration, and explainable AI within a unified framework. Taking inspiration from it, our paper introduces a novel hybrid architecture, named X-MHD, which incorporates DistilBERT representations, knowledge graph embeddings, and explainable artificial intelligence to achieve interpretable mental health classification which consists multiple classes using text taken from social media. The next section overviews the architecture and components of the proposed X-MHD approach.

III. METHODOLOGY

We introduce X-MHD, a unified and efficient approach that can detect emotional well-being issues in internet communities with high accuracy and transparency. Our approach combines three components: contextual language understanding using a DistilBERT encoder, semantic augmentation using representations from external knowledge graphs, and explainable artificial intelligence techniques, which are collectively represented as "X" within X-MHD. Our approach's processing pipeline involves text preprocessing, knowledge graph-enhanced contextual encoding, feature fusion, multi-class prediction, and interpretability. As shown in Figure 1, X-MHD's efficient design using a lightweight transformer structure along with structured semantic knowledge and interpretability results in

a favorable trade-off between prediction accuracy, efficiency, and interpretability.

A. Motivation and Design Considerations

Mental health issues expressed on social media are often subtle and context-dependent. These issues are difficult to detect using traditional text classification models that cannot capture implicit psychological cues and contextual clues present within social media text. To overcome these challenges, we introduce a novel X-MHD approach that incorporates contextual language representations using DistilBERT, structured semantic knowledge using knowledge graphs, and explainable AI techniques to achieve accurate mental health predictions.

B. Data Collection

Experiments have been performed using the Reddit dataset, called the Reddit Suicide Watch and Mental Health (SWMH), which consists of a number of mental health-focused subreddits from the Reddit platform [19]. The dataset primarily consists of posts related to issues like depression, anxiety, and bipolar disorder. These are all mental health based. The dataset contains a total of 54,412 labeled posts, which makes it a complete dataset for mental health classification tasks. The dataset has five tags: Anxiety, bipolar, depression, SuicideWatch, and offmychest. Also all tags having prefix "Self.". It is divided into a data to train, data for validation, and data to test in a ratio of 70:10:20. The labels have been remapped to Anxiety, Bipolar, Depression, Suicide, and None, respectively, for better interpretation of the results. The OffMyChest category has been included as a neutral category, which does not indicate any mental health disorder. The posts included in this category express emotional content but do not indicate any particular mental health disorder.

C. Data Processing

The social media texts on mental health issues might contain informal writing, irrelevant information, and contradictions. To guarantee high-quality input, the following preprocessing steps are taken:

- Noise removal: removal of URLs, HTML tags, mentions, hashtags, and other non-text symbols
- Normalization: converting all input characters to lower-case and limiting excessive character repetition (e.g., replacing "soooo" with standard whitespace normalization)
- Text standardization: unicode normalization and stop word filtering, as appropriate
- Label encoding: applying a multi-class encoding scheme on mental health categories, which includes:
 - Depression
 - Suicide Watch
 - Disorder Off (Note: The category names can be modified according to the exact categorization scheme used)

D. Knowledge Graph Construction and Integration

To enrich contextual representations with structured semantic knowledge, we incorporate a domain-specific knowledge graph (KG) representing relationships among mental health concepts. The knowledge graph consists of entities such as symptoms, emotional expressions, psychological states, and mental health conditions extracted from mental health lexicons and Reddit mental health discussions.

Each entity in the knowledge graph is represented using the TransE embedding method, which learns low-dimensional vector representations by modeling relationships between entities as translation operations in the embedding space. Formally, the embedding model optimizes the relationship:

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t} \quad (1)$$

The head and tail entity embeddings are denoted by $\mathbf{h}$ and $\mathbf{t}$, respectively, and $\mathbf{r}$ represents the relation embedding vector.

The resulting knowledge graph embeddings provide semantic representations of mental health concepts that complement the contextual embeddings produced by DistilBERT. These embeddings are then integrated with textual representations during feature fusion by enhancing the capturing ability of model implicit psychological signals. The contextual representation obtained from DistilBERT and the knowledge graph embedding are fused using feature concatenation:

$$\mathbf{h}_{\text{fused}} = [\mathbf{h}_{\text{text}}; \mathbf{h}_{\text{KG}}] \quad (2)$$

where $\mathbf{h}_{\text{text}}$ represents contextual embeddings generated by DistilBERT and $\mathbf{h}_{\text{KG}}$ represents knowledge graph embeddings. The fused representation is then passed to the classification layer for the final prediction.

E. Large Language Models

Emotional well-beings prediction problems are evaluated by a number of text-based (language) models to check effectiveness

- BERT: This model was fine-tuned to apply the model to the SWMH dataset.
- RoBERTa: This is an optimized version of the BERT model. It was fine-tuned to enable performance comparisons with the baseline model.
- MentalBERT and MentalRoBERTa: These are domain-specific transformer models that were trained on text data related to mental health. They were used directly.
- MentaLLaMA: This is a language model based on the LLaMA2 model. Evaluation was done in predicting mental health outcomes.
- BART-Large-MNLI: Zero-shot classification was used to predict mental health classes independently.

F. Evaluation Metrics

For evaluating the model comprehensively, various metrics were used:

- Accuracy: This metric calculates the overall accuracy of predictions made by the model.

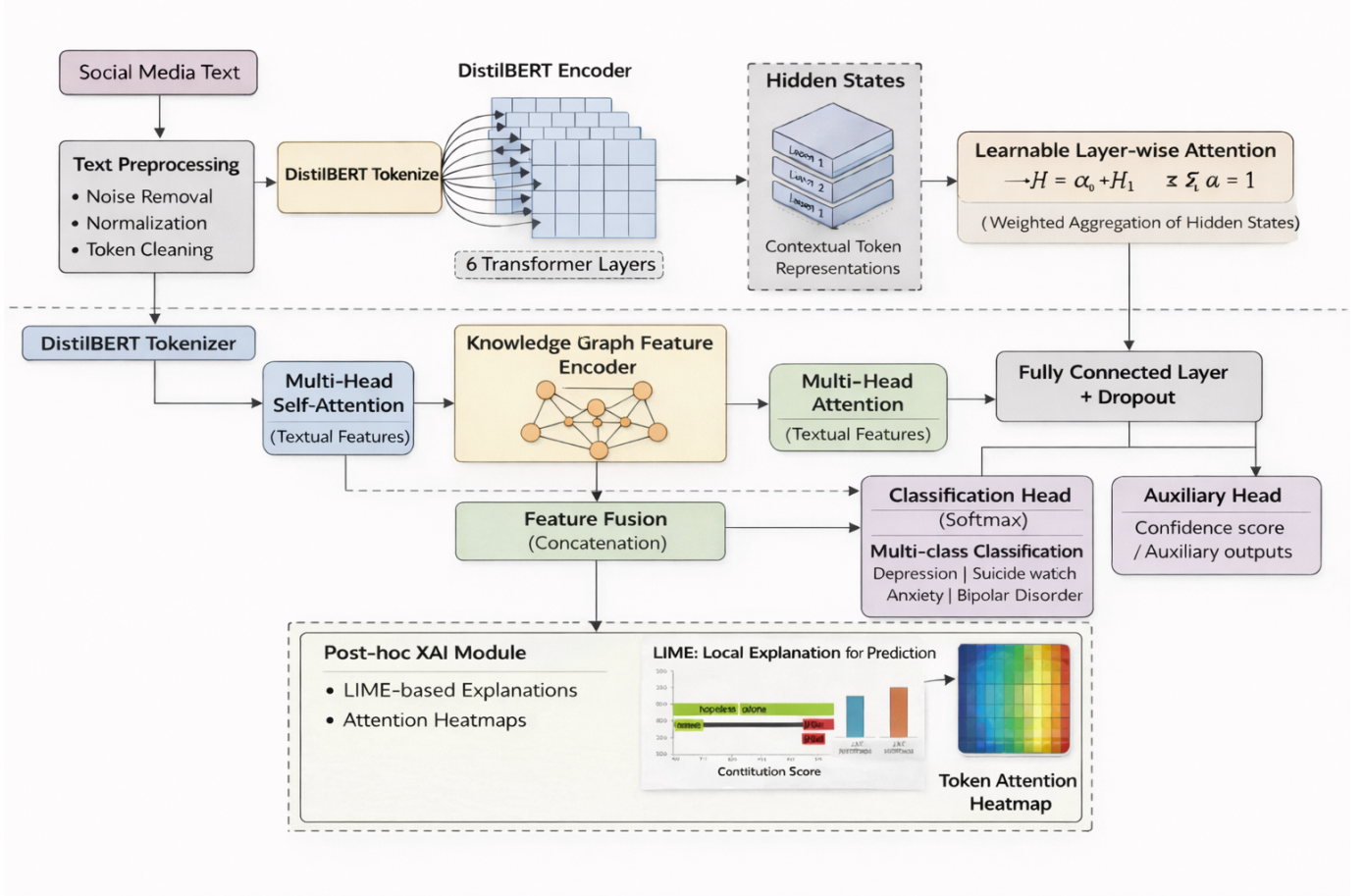


Fig. 1: Overall architecture of the proposed X-MHD model integrating DistilBERT, Knowledge Graph features, and LIME-based explainable AI for multi-class mental health disorder detection.

- **Classification Report:** This metric provides detailed information regarding precision, recall, and F1 score for different classes.
- **Confusion Matrix:** Calculation of true positives, true negatives, false positives, and false negatives is done. These are needed to create confusion matrix.

G. Multi-Task Learning Framework

To improve model robustness and generalization, we adopt a multi-task learning framework consisting of a primary classification task and an auxiliary prediction task. The primary task performs multi-class classification of mental health conditions, while the auxiliary task predicts a confidence score indicating whether the input text expresses mental health related distress.

The overall training objective is defined as a weighted combination of the two loss functions:

$$L_{\text{total}} = L_{\text{main}} + \lambda L_{\text{aux}} \quad (3)$$

where L_{main} represents the multi-class classification loss and L_{aux} denotes the auxiliary task loss. The parameter λ controls the contribution of the auxiliary task during training.

H. Hyperparameter Settings

The proposed X-MHD model employs *DistilBERT-base-uncased* as the pretrained language encoder, consisting of six transformer layers. Maximum sequence length of 256 is used to tokenize input texts. Batch size of 32 and optimized with the AdamW optimizer at a learning rate of 1×10^{-5} . To reduce overfitting, a dropout rate of 0.3 was applied to fully connected layers. The multi-head attention module utilizes eight attention heads, while the knowledge graph encoder generates embeddings of dimension 50. A class-weighted cross-entropy loss is adopted to address class imbalance, complemented by an auxiliary loss weighted at 0.3 to improve training stability. The model is trained for 10 epochs with 100 warmup steps, and performance's evaluation is done using 5-fold cross-validation, with all hyperparameters kept fixed across folds to ensure reproducibility.

IV. EXPERIMENTAL RESULTS

A. Ablation Study

To understand the contribution of individual components in the proposed architecture, we perform an ablation study by

incrementally removing different modules of the model. The results are shown in Table I.

TABLE I: Ablation Study of X-MHD

Model Variant	Accuracy
DistilBERT only	78%
DistilBERT + Knowledge Graph	83%
DistilBERT + XAI	80%
DistilBERT + KG + XAI	85%
Full X-MHD (Proposed)	88%

We measured the performance of our model using accuracy, precision, recall, and F1-score. as shown in Table I, fine-tuned BERT-base and RoBERTa achieved accuracies of 71% and 72%, respectively, while domain-adapted models such as MentalBERT and Men- The RoBERTa model achieved a performance of up to 70%. In contrast, large and zero-shot models, such as BART-large MNLI and MentalLLaMA, comparatively lower accuracy, emphasizing their relative ineffectiveness in this task without task-specific adaptation. The proposed X-MHD model, developed using a DistilBERT-based contextual encoder enriched with knowledge graph representations, achieved the highest test accuracy of 88% while maintaining a parameter-efficient architecture. As shown in Fig. 2, the LIME explanation highlights emotionally significant tokens such as “suicidal” and “wont”. While less relevant terms prevent negative influence, exemplifying the interpretability of the proposed model. As shown in Fig. 3, the model shows excellent class-wise discrimination, with minor confusion among semantically overlapping categories. We conducted a paired t -test across the 5-fold cross-validation results to check performance of our model. The improvements achieved by the proposed X-MHD model over baseline models were found to be statistically significant ($p < 0.05$).

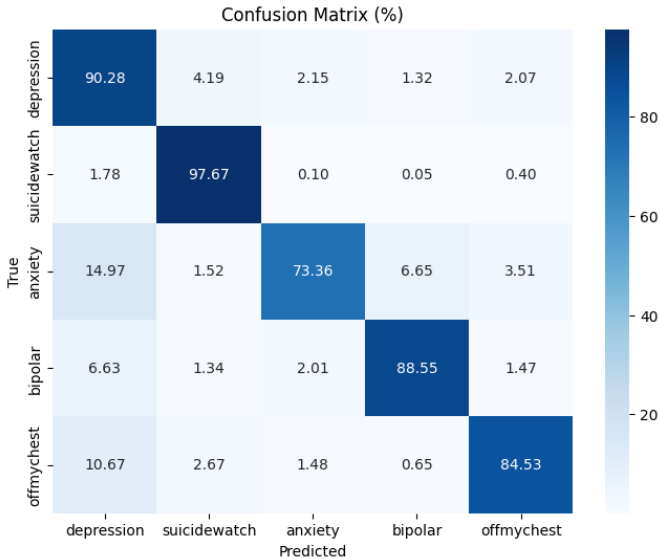


Fig. 2: F1-score comparison across mental health conditions.

Although other explainability methods such as SHAP and Integrated Gradients have been explored in prior work, LIME

TABLE II: Summary of Experimental Results on the Test Dataset

Model	Parameters	Accuracy
MentalBERT-base	110 M	69%
MentalRoBERTa	110 M	70%
MentalLLaMA-7B	7 B	59%
<i>Fine-Tuning</i>		
BERT-base	110 M	71%
RoBERTa-base	110 M	72%
<i>Zero-Shot Prompting</i>		
BART-large-mnli	400 M	49%
<i>Proposed Model</i>		
X-MHD	60 M	88%

TABLE III: Per-Class Performance of X-MHD

Class	Precision	Recall	F1-Score
Depression	0.89	0.87	0.88
Anxiety	0.85	0.83	0.84
Bipolar	0.86	0.85	0.85
SuicideWatch	0.90	0.91	0.90
OffMyChest	0.82	0.84	0.83

was selected in our framework as it provides clear local explanations for transformer-based models.

V. ETHICAL CONSIDERATIONS

This study uses publicly available Reddit data from mental health communities. Although anonymized, ethical concerns regarding privacy, bias, and responsible deployment must be considered. Suggested model is intended for research purposes and not for clinical diagnosis.

VI. LIMITATIONS

A community, Reddit is the source from where that dataset is taken. This means that there could be some level of demographic bias. Moreover, using LIME or attention mechanisms to provide explanations may not be able to capture the intricacies of neural networks. Therefore, this system should be used as an assistant tool rather than a diagnostic tool.

VII. CONCLUSION

The current study introduces X-MHD, which is an efficient and interpretable approach to conducting multi-class classifications on human well-being states using internet community



Fig. 3: LIME-Based Local Explanations for Token-Level Contributions to Predicted Mental Health Category.

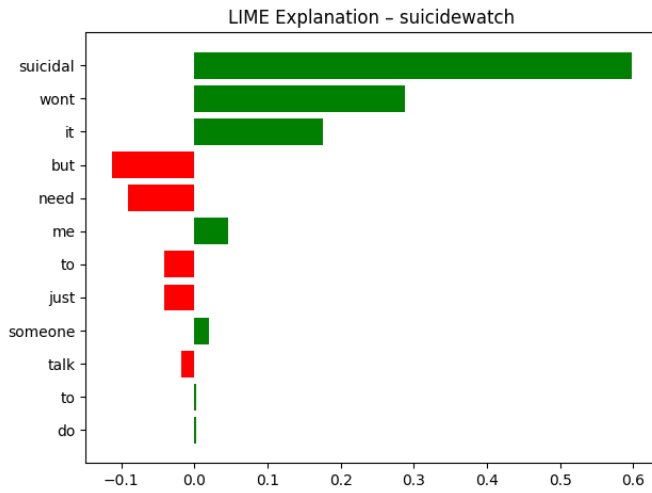


Fig. 4: LIME-Based Local Explanation for Suicide Watch Prediction, with Token-Level Contributions.

text. This approach combines the results of using DistilBERT for contextual representations, knowledge graph embeddings, and explainable artificial intelligence. X-MHD achieves higher accuracy compared to various transformer models, which is shown in the result section. This shows that using contextual representations along with semantic knowledge is essential in analysis of mental health.

REFERENCES

- [1] World Health Organization, "Mental disorders," <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>, 2024, accessed: Nov. 2024.
- [2] D. D. Ebert, C. Buntrock, P. Mortier, R. Auerbach, K. K. Weisel, R. C. Kessler, P. Cuijpers, and J. G. Green, "Prediction of major depressive disorder onset in college students," *Depression and Anxiety*, vol. 36, pp. 294–304, 2019.
- [3] B. Kaushik, A. Sharma, A. Chadha, and R. Sharma, "Machine learning model for sentiment analysis on mental health issues," in *2023 15th International Conference on Computer and Automation Engineering (ICCAE)*. IEEE, 2023, pp. 21–25.
- [4] L. D'Cruz, V. Dubey, and P. Thakur, "Depression prediction from combined reddit and twitter data using machine learning," in *2023 2nd International Conference for Innovation in Technology (INOCON)*. IEEE, 2023, pp. 1–5.
- [5] J. Q. Chen, K. Qi, A. Zhang, M. Shalaginov, and T. H. Zeng, "Covid-19 impact on mental health analysis based on reddit comments," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 2253–2258.
- [6] C. Dwivedi and J. Rao, "Multilingual sentiment analysis for detecting mental health problems using a hybrid algorithm combining rnn and bi- lstm," in *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2023, pp. 527–534.
- [7] S. Sharma, G. S. P. Ghantasala, V. Gupta, P. K. Sarangi, R. S. R., and A. Sungeetha, "Analyzing employees' mental health using machine learning algorithms," in *2025 12th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2025, pp. 1–5.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, E. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [9] T. Bikku, J. Jarugula, L. Kongala, N. D. Tummala, and N. V. Donthiboina, "Exploring the effectiveness of bert for sentiment analysis on large-scale social media data," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2023, pp. 1–4.
- [10] A. Joshy and S. Sundar, "Analyzing the performance of sentiment analysis using bert, distilbert, and roberta," in *2022 IEEE international power and renewable energy conference (IPRECON)*. IEEE, 2022, pp. 1–6.
- [11] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," in *proceedings of the thirteenth language resources and evaluation conference*, 2022, pp. 7184–7190.
- [12] V. Vajre, M. Naylor, U. Kamath, and A. Shehu, "Psychbert: a mental health language model for social media mental health behavioral analysis," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1077–1082.
- [13] C. H. R. Madhuri, J. Sankar Krishna Bandaru, M. Srinu, and G. M. Anand Vardhan, "Ai-powered mental health screening and support for homeless children," in *2025 AI-Driven Smart Healthcare for Society 5.0*, 2025, pp. 115–120.
- [14] S. Kumar, A. R. Chauhan, Akhil, A. Kumar, and G. Yang, "Respboostnet: Mental stress detection from biomarkers measurable by smart-watches using boosting neural network technique," *IEEE Access*, vol. 12, pp. 149 861–149 874, 2024.
- [15] A. R. Sarode, M. V. Rani, B. Das, U. S., and S. Punia, "An ai-based framework for interpretable mental health literacy segmentation and decision support," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 12, pp. 8799–8806, 2025.
- [16] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "Mentallama: interpretable mental health analysis on social media with large language models," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 4489–4500.
- [17] P. R.S., H. Deep, and D. Yadav, "Mental health support using gen-ai shot prompting technique and vector embeddings," in *2025 International Conference on Data Science and Business Systems (ICDSBS)*, 2025, pp. 1–5.
- [18] H. K. Marrapu, B. Maram, and P. Reddi, "New analytic framework of public mental health prediction using data science," in *2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, 2022, pp. 1–6.
- [19] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10 695–10 707, 2021.

SarcasmBERTa: Multilingual Sarcasm Detection using RoBERTa and multi head attention based model

SUMAN SAHA, Computer Science and Engineering, Delhi Technological University, Delhi, India, India

SANJAY KUMAR, Computer Science and Engineering, Delhi Technological University, Delhi, India, India

In language analysis, detecting sarcasm is vital for correctly interpreting users' intent and sentiment. The content posted on various social media platforms is multilingual and includes low-resource languages such as Hindi, Bengali, etc. We propose a novel model named SarcasmBERTa using the Robust Optimized BERT Approach (RoBERTa) with multi-head attention (MHA) layers, which enables it to jointly capture a variety of semantic clues in the input to achieve effective sarcasm identification in multi-lingual and low-resource languages. In our model, we used the RoBERTa approach that tokenizes the input text using Byte-Pair Encoding (BPE), where it splits the input words into subword units if needed. Then we fine-tuned the model using MHA and layer normalizations. Weighted cross-entropy loss is used to handle class imbalance. It generates context-aware embeddings, meaning that the same word can have different embeddings based on its context. To prevent overfitting, cosine learning rate scheduling with dropout regularization and warmup is employed, which ensures the proposed model trains more robustly and smooth gradient flow during early epochs. We conduct experiments on open, heterogeneous, and multilingual datasets to evaluate the model's efficiency. The proposed model achieves an average accuracy of 96.10 and an average F1 score of 96.15 in percentage on the evaluated datasets.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Multilingual and cross-lingual methods*; • **Artificial intelligence** → **Machine learning**; *Deep learning*.

Additional Key Words and Phrases: Class Imbalance, Multi-Head attention, Multilingual Sarcasm Detection, Robust Optimized BERT Approach (RoBERTa), Low-Resource Languages

ACM Reference Format:

Suman Saha and Sanjay Kumar. 2026. SarcasmBERTa: Multilingual Sarcasm Detection using RoBERTa and multi head attention based model . 1, 1 (May 2026), 27 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Authors' Contact Information: Suman Saha, Computer Science and Engineering, Delhi Technological University, Delhi, India, Delhi, India, sumansaha0178@gmail.com; Sanjay Kumar, Computer Science and Engineering, Delhi Technological University, Delhi, India, Delhi, India, sanjay.kumar@dtu.ac.in.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2026/5-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Sarcasm mixes humor and irony to point out inconsistencies, and understanding it helps us to get the real meaning in conversations. Posted contents on internet communities sometimes may be multilingual. It is important to figure out when someone is being sarcastic, helping us to understand what actually they really mean or feel. For generations, sarcasm is an aspect of human communication as a subtle yet powerful form of expression. Sarcasm, which has its roots in ancient languages and literature, enables people to express ideas that are tinged with comedy, irony, or subtly critical remarks. As defined by the Macmillan English Dictionary, sarcasm involves saying or writing the opposite of one's intended meaning, frequently used to criticize or mock particular viewpoints [17]. In the current digital era, sarcasm is highly prevalent in online interactions, like on social networks, forums, and news portals. Detecting sarcasm is hard, especially for automated systems such as sentiment analysis. Misinterpretation of sarcasm happens due to a contradiction between the literal and intended meaning. We can understand it through an example: "Wow, you are early!" This quote may show positivity, but in a sarcastic context, it reflects another thinking or disappointment. To refine these challenges, sarcasm detection is essential for improving natural language understanding (NLU) models [57]. Sarcasm detection in text [25] is difficult as it depends on context, cultural differences, and subtle linguistic nuances. To understand the literal meaning behind the statement or text, traditional text analysis tools are used. Sarcasm usually involves wordplay, exaggeration, or contradictions that can confuse even advanced systems. As an example, the phrase "What a genius plan!" may show sincerity or sarcasm, depending on the situation we are in. If we don't get extra content about the phrase, it is hard to tell the difference. In spoken conversations [16], sarcasm is often supported by tone of voice, facial expressions, or gestures, which are completely missing in written text. This lack of visual or auditory cues makes sarcasm detection in text rely heavily on surrounding information [49], like the rest of the conversation or the speaker's usual style. Identifying sarcasm is as difficult for humans as well as for computers [41]. Sarcasm appears differently in various languages and cultures. In the earlier stage, researchers apply supervised learning techniques by the use of n-gram features, POS tags [1], and sentiment-based attributes. SVMs and decision trees are common but require intensive [37] feature engineering and do not generalize well across domains. After RNNs and CNNs other transformed models like BERT, GPT, and XLnet come to the field and understand sequences better [15]. LSTM and BiLSTM architectures enable the capturing power of long-range dependencies [43] in sarcastic statements. Multilingual context gets its popularity, so models like mBERT and XLM-R [14] become popular. Simultaneously, multimodal sarcasm detection using image-text or audio-text data gained traction to incorporate non-verbal cues. While English speakers might use exaggeration or irony, other languages might use idioms or unique sentence structures. Traditional models, which are only learned on one single language, which is challenging to multilingual scenarios and this is the most vital challenge in sarcasm detection in different linguistic groups. However, the labeling of sarcasm is subjective in nature, and different people

might have different views and interpretations of the same thing.

Joshi et al. [26] define the process of detecting sarcasm as a cognitive and computational process. Due to the anticipated sentiments, detecting sarcasm might be challenging. As we can take an example, the sentence 'There's nothing I enjoy more than sitting in traffic for hours!' would be expected to have a positive sentiment, whereas in reality, it is negative. Traditional sentiment analysis tools would fail to detect such sentences, leading to incorrect results. Accurate detection of sarcasm is important for various purposes, such as understanding opinions, [23] analysis of sentiments, and even understanding mental states. Sarcasm detection models, however, fail to perform their expected functions, especially in the case of multiple languages. For instance, there is a scarcity of multilingual sarcasm datasets, which is a significant issue. The majority of the sarcasm dataset available today is in English, [21] which has hindered the development of models that would contextualize sarcasm and understand it. Moreover, the complexity of sarcasm, which defines it, is an issue that would make it extremely difficult for machines to detect sarcasm. These challenges motivate us to develop a novel framework to detect sarcasm that would overcome the challenges of sarcasm detection models. If we can check the effectiveness of sarcasm detection, we can increase knowledge about sarcasm in human languages. Potentially enable smarter, more effective natural language processing systems for human communication in all its nuances. RoBERTa, which is based on a deep understanding of language. The creation of an improved understanding of sarcastic expressions is made possible by the capacity of RoBERTa to comprehend the complicated relationships between words and phrases, as well as the application of the attention mechanism consist of multiple heads. This is more impactful when the datasets are multilingual, given the global nature of digital communication. This aspect of the application, which is multilingual, addresses the gap in the literature since most research is based on a singular language, namely English, and ignores the linguistic and cultural diversity that influences the use of sarcasm. SarcasmBERTa is effective and flexible, capable of handling the vast variety of language expressions and structures within the multilingual datasets. It is also capable of global applications, including sentiment analysis for businesses operating globally, social media content moderation, and customer feedback analysis, among others. We discuss the possibility of using the RoBERTa model in conjunction with MHA mechanisms. The integration of MHA mechanisms with the RoBERTa model makes it more accurate and reliable in detecting sarcasm. Contribution of this paper is described below.

- An improved model is proposed for sarcasm detection in multi-lingual environments, including low-resource languages, using a pre-trained model named RoBERTa with MHA mechanisms for complex contextual relationships in text data.
- A loss function is used, weighted cross-entropy for class balancing, where more weight is given to the minority class (sarcasm) in our model.
- The paper proposes an effective training method using techniques such as k-fold cross-validation, cosine learning rate scheduling, dropout, and warm-ups,

which work in conjunction with each other to improve training stability, prevent overfitting, and guarantee reliable performance.

- Achieved a novel improvement compared to other models and existing sarcasm detection techniques. This is in both English and low-resource languages. This is an improvement in sarcasm detection in multi-lingual environments.

Formation of the paper is shown in the following sections. Recent advances in sarcasm detection techniques are discussed in Section 2. The different steps of our proposed framework are explained in Section 3. Section 4 describes the data sets and the evaluation metric we used for our datasets taken. The result is shown in Section 5, which we obtained from our experiments. The conclusion of the paper shows the future, what are the key insights and suggesting direction for future research shown in section 6.

2 Related Work

In recent years, sarcasm detection became an essential task in Natural Language Processing (NLP) and multimodal analysis. Here we introduced related works in baseline, deep learning, transformer-based, and also in multimodal fusion works.

2.1 Baseline and Deep Learning-Based Models

Ahuja and Sharma [3] propose the LMTweets encoder model for sarcasm and irony detection, trained on 500,000 tweets. Cambria et al. [35] proposed a multi-task learning framework based on a deep neural network to model the connection between sentiment and sarcasm while simultaneously solving both tasks. Cai et al. [9] developed a multi-modal sarcasm detection model for Twitter that combined text, images, and image attributes. Their hierarchical fusion model first extracted features from each modality and then integrated them using a bidirectional LSTM network. Zhang et al. [54] created ConAttSD, a model designed to detect sarcasm in multimodal situations, like video conversations. Use of a contrastive attention mechanism, which highlights the differences between positive text paired with a negative tone. Collection of 1.3 million sarcastic statements from Reddit that were self-annotated, which was the largest sarcasm dataset done by Khodak et al. [29]. Greeshma and Simon [19] proposed a movie review classification system using a Bidirectional Gated Recurrent Unit (BiGRU) combined with a custom attention mechanism and GloVe embeddings.

Rosid et al. [48] developed a hybrid model called MHA-CovBi for detecting sarcasm in Indonesian-English mixed-code texts. The model combined CNN with multi-head attention and BiGRU, using FastText for Indonesian and GloVe for English word embeddings. Pandey and Singh [40] proposed a sarcasm detection model by combining LSTM and BERT for code-mixed language. Yue et al. [53] developed KnowleNet, a multimodal sarcasm detection model that used ConceptNet to incorporate prior knowledge and improve understanding of image-text by detecting cross-modal semantic similarities. A new hybrid model was created by Hazarika et al. [22] that tried to detect sarcasm by analyzing authors' past behavior in comments and studying the general behavior of authors within specific subreddits. Bao et al. [6] introduced the HIAN

model to address the challenge of aligning diverse sarcasm cues from text and images. It utilized a BiLSTM with interactive attention to extracting richer textual features by integrating visual class words.

2.2 Transformer and Pretrained Language Model (PLM)-Based Models

A new approach for sarcasm detection that used context from parent comments on Reddit is proposed by Yang et al. [52]. They worked with the SARC dataset and employed a pre-trained XLNet model to extract features. An approach for detecting sarcasm in Twitter posts using the transformer model has been proposed by Gregory et al. in [20]. The authors also used LSTM, GRU, and transformer models. The top-performing model in this approach is an ensemble of transformer models, including BERT, RoBERTa [33], XLNet [52], RoBERTa-large, and ALBERT [31]. At some points in the text, an emoji is used, followed by how it is interpreted by the model, and then how it is used in a model which incorporates it SEEmoji MUSTARD [12], an extended dataset which includes emojis and uses them for sentiment and emotion annotations for better sarcasm detection in multimodal conversations. MUSTARD dataset [11] is utilized for multimodal-based sarcasm detection. The dataset is based on TV shows, which consist of audio-visual dialogues with sarcasm labels and the dialogue history. Internet community sites such as Reddit and Twitter are highly populated with figurative language. Nikhil Jaiswal [27] suggested a deep neural structure for the detection of sarcasm. He utilized pre-trained language models such as BERT and RoBERTa. He then used them to improve their performance on a Twitter dataset. A sarcasm detection model is proposed by Meng et al. [36], combining BERT and an attention mechanism to address the challenge of detecting sarcasm in sentences with complex semantics and few sentiment words. However, sarcasm's implicit nature makes it challenging to detect, [4], particularly in Bangla, where research remains limited compared to English. Anan et al. introduced a BERT-based model, which was much better than traditional data-driven techniques. To test their approach, they created BanglaSarc, a new Bangla sarcasm dataset made up of comments from Facebook and YouTube. HumourHindiNet, a framework for detecting humor in Hindi web series, was created by Kumar et al. [30]. They built a new multimodal dataset using dialogues from shows like 'Kota-Factory' and 'Panchayat', which were manually labeled as humor or non-humor. Pramanick et al. [44] introduced a new framework called MuLOT (Multimodal Learning using Optimal Transport). This framework has been tested using benchmark datasets such as MUSTARD, UR-FUNNY, and MST. [10]. Combining textual linguistic features with visual information, the multimodal sarcasm detection approach was proposed by Han et al. [9]. Their model focused on the textual modality, extracting high-frequency characters from pure text and hashtags to create a third modality alongside images and text. Qin et al. [45] created MMSD 2.0, an improved version of the original MMSD dataset, to fix problems like misleading cues and unrealistic negative sample. Sometimes language comes into the problem. Bedi et al. [7] developed MaSaC, a Hindi-English code-mixed dataset for sarcasm detection and humor classification in conversations, and proposed MSH-COMICS, an attention-rich neural model. Liang et al. [32] introduced an approach to address the challenge

of identifying sarcasm using both text and images. The method employed object detection and OCR to extract key visual and textual cues, which were modeled using multimodal graphs. A graph-based contrastive learning strategy was proposed to enhance modality interaction and representation. In multimodal, we need to combine text, images, and image attributes. A huge corpus of Bengali text for the specific problem of research on sarcasm detection in Bengali, named ‘Ben-Sarc,’ comprising a total of 25,636 comments gathered manually from various Facebook public pages and rated by external evaluators, was proposed by Lora et al. [34]. Pokhriyal and Jain [42] proposed a sarcasm detection approach for Hindi social media comments using a probabilistic non-linear optimization technique integrated with word embeddings. For the Algerian dialect, Bousmaha et al. [8] introduced an integrated multimodal approach to sarcasm detection by utilizing both visual cues through VGG-19 and text cues through EasyOCR with some linguistic properties and machine learning algorithms. Zhang et al. [55] proposed a Multimodal Interaction Learning (MIL) framework for joint sarcasm and sentiment detection that employs cross-modal target attention and a dual-gating network to capture shared and task-specific features across modalities. Rodríguez et al. [47] proposed SINCERE, a hybrid graph-based sarcasm detection framework that integrates sentiment and emotion analysis using distant supervision and graph-based latent representation learning. Zhang et al. [56] introduced AMSD Dataset and the SarcasmBench benchmark to evaluate sarcasm understanding across 16 LLMs. While previous models performed well on certain datasets, they still had limitations in multilingual and low-resource settings. By addressing this issue we proposed a model that overcame these limitations and achieved good performance. Refai et al.[46] proposed a multi-dimensional prompt evaluation framework for LLMs that assesses prompts based on similarity, performance, efficiency, and consistency across several Arabic NLP tasks—including sarcasm detection—using models such as GPT-4o mini, LLaMA, and Claude 3.5 Haiku, highlighting trade-offs between prompt effectiveness and computational efficiency. Bade et al. [5] investigated pragmatic generalization of LLMs for multilingual sarcasm detection by fine-tuning encoder models such as BERT, RoBERTa, XLM-RoBERTa, and DistilBERT, while comparing them with few-shot GPT-4o, showing that fine-tuned models significantly outperform generative LLMs in cross-lingual sarcasm detection tasks. According to Ou & Li et al. [39], a multi-granularity information fusion network can be used for the detection of multimodal sarcasm using multi-task CLIP and global-local cross-modal interaction learning. Various machine learning and deep learning architectures, including Random Forest, XGBClassifier, fastText-stackedTCN, and mBERT, were explored by Ahmad et al.[2] on a dataset of 1,945 humorous Hindi newspaper headlines from Hindustan and Dainik Jagran. Although previous models achieved success on certain datasets, they are still not sufficient for low-resource and multilingual scenarios. We proposed a model that transcends these limitations and achieves success on these problems.

3 Methodology

The suggested paradigm for the identification of sarcasm, which utilizes the RoBERTa model and the multi-head attention mechanism, is described in this section. We start

our proposed work with transforming textual input into vectorized inputs via the RoBERTa tokenizer, which maintains the contextual essence of each token. These embeddings are fed into a pretrained RoBERTa model, which is good at identifying subtle semantic patterns, including sarcasm. Multi-head attention layers are included, enhancing the model's focus on critical textual cues to refine sarcasm detection. We employ weight loss functions to ensure balanced learning across classes. For the reliability and consistency of the suggested model, the K-fold cross-validation approach is utilized and subsequent section highlights the various stages involved in the proposed model.

3.1 Data Preprocessing

We start with data that is already public, although it is not clean. It is raw and contains clutter like URLs and image IDs. The data used in this work is from KnowleNet, which needs to be cleaned. For 'Ben-Sarc,' the labels are already provided as "sarcasm" and "not sarcasm." For our work, we standardize the labels by using 1 for "sarcasm" and 0 for "not sarcasm." We use the same labels for all the data. While cleaning the data, we realized that there are duplicate data in some of the data sets, such as 'Ben-Sarc' and Hindi. For ease, we also clean the paragraph formatting in the Hindi data. For k-fold validation, we use all the validation, test, and training data in one dataset for each language, such as English, Hindi, and Bengali. Our method helps in making the process of preprocessing and validation in various data sets efficient.

3.2 Tokenization: Preprocessing Text for Model Input

Roberta tokenizer is used in this paper. This tokenizer is important in facilitating the transformation of the unstructured input into an understandable language for the model. This tokenizer is initialized with the weight parameters of the 'roberta-base' model from Hugging Face. To ensure standardization in input data, truncation, and padding were used where necessary.

- Truncation: Inputs with more than 256 tokens were truncated to the defined max limit, emphasizing key content.
- Padding: Tokens were used for shorter input texts to ensure consistency in data format.

For each input text, tokenization produces two outputs: attention masks and token ID outputs. The model is capable of processing text data in numerical form, token ID, which is an integer value allocated to each token. The attention masks represent valid token positions (1) and padding positions (0). The tokenizer processes inputs in batches, thereby minimizing time delay in data flow.

3.3 Model Design and Training Approach

The enhanced model structure for the sarcasm identification challenge is based on the RoBERTa model. The multi-head attention method (Fig. 1) is used to enhance the understanding of the context in the text parts. Avoiding overfit is important things so layer Norm and dropout rate 0.3 are used. Which helps in stability during the training

process. The output of the RoBERTa model is classified using a classifier. The weighted loss function [0.7, 1.3] is used to enhance the model to take care of class balance in the dataset. The optimization process is carried out using the AdamW optimizer with a learning rate of $1e-5$. For the fine-tuning of the training process, the cosine annealing schedule with warm restarts is used. This schedule gradually decelerates the learning process and restarts it at regular intervals. This helps the model avoid the local optima and converge better. The above-discussed strategies are the most optimized and balanced approach for the problem of sarcasm detection. The proposed SarcasmBERTa model process is illustrated in Figure. 2. Raw data collection marks the first step in the process. Raw data will then be subjected to the data preprocessing step where it undergoes several procedures such as data cleaning through the process of lemmatization, stemming, replacement of emojis with texts, among others. After going through the data preprocessing stage, the data will then be tokenized using the RoBERTa (Fig. 1) model's tokenizer. This model employs the K-Fold Cross-Validation technique for training, and the tokenized input data will be replicated five times.

3.3.1 Tokenizing the Input Text. Let's take an input text T , to process it, we first tokenize it using a pre-trained tokenizer which is the RoBERTa tokenizer. The tokenizer converts the text into a sequence of integer token IDs, ensuring a fixed-length L using padding or truncation. This fixed-length token sequence becomes the input to the model. Here, X is a function of T , representing a sequence of token IDs produced by the tokenizer.

$$X = \text{Tokenizer}(T) = \{x_1, x_2, \dots, x_L\}, \quad x_i \in \mathbb{Z} \quad (1)$$

where x_i represents the token ID for the i -th token in the sequence.

3.3.2 Generating Contextual Embeddings with RoBERTa. As we get the tokenized input, X it is passed through the pre-trained RoBERTa model. It produces contextual embeddings for each token, incorporating the context of the surrounding tokens. These embeddings form a matrix H representing the entire input sequence.

$$H = \text{RoBERTa}(X, M) = \{h_1, h_2, \dots, h_L\}, \quad h_i \in \mathbb{R}^d \quad (2)$$

Here $H \in \mathbb{R}^{L \times d}$ is the hidden state matrix. $M \in \{0, 1\}^L$ is the attention mask indicating valid (non-padded) tokens. The RoBERTa hidden layer has a dimensionality of d .

3.3.3 Refining Representations via Multihead Self-Attention. While the conventional attention mechanism in the RoBERTa transformer architecture is already employed, an additional MHA layer is added on top of the context-aware representations to enhance token-level interactions specific to sarcasm contexts.

Multihead attention is a good approach for contextual understanding in text. So, to further enhance the contextual representations, a multihead self-attention mechanism is applied. For each attention head, we calculate the attention output as:

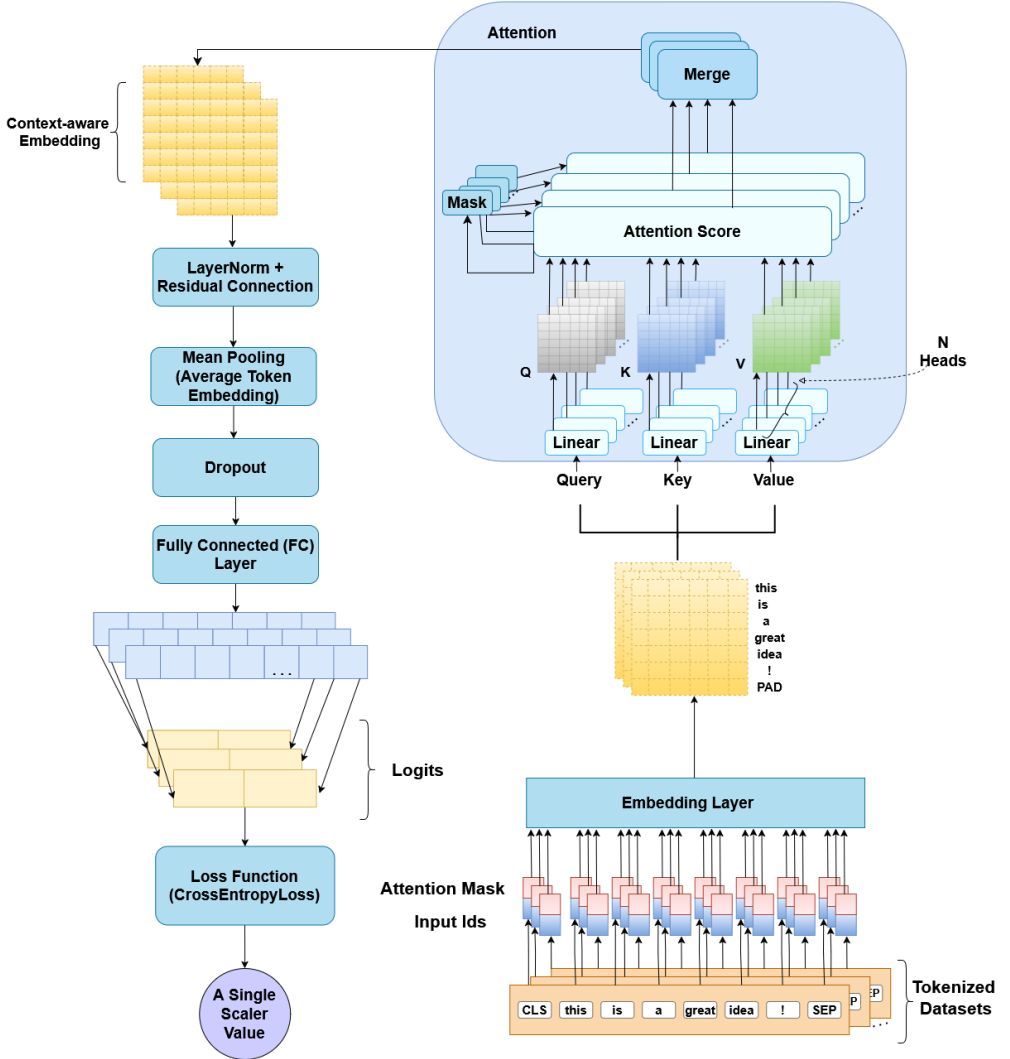


Fig. 1. Architecture of the proposed sarcasm detection SarcasmBERTa model

$$\text{Attention}_k(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3)$$

where $Q = HW_Q$, $K = HW_K$, $V = HW_V$ query, key and value vectors respectively. $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learned projection matrices and $d_k = d/\text{num_heads}$ is the dimensionality of each head. All attention head outputs are concatenated, which

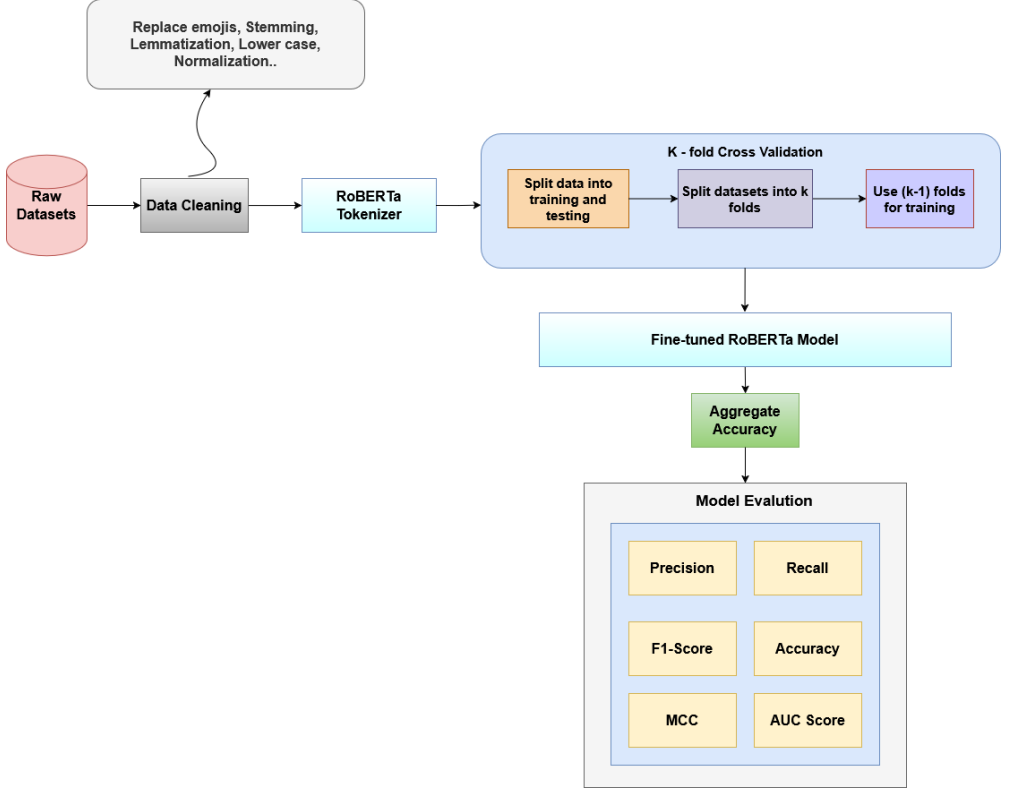


Fig. 2. Workflow of the Sarcasm Detection Model, i.e., SarcasmBERTa

allows the model to project the resulting vector back into the original embedding space.

$$H' = \text{Concat}(\text{Attention}_1, \dots, \text{Attention}_k)W_O \quad (4)$$

where $W_O \in \mathbb{R}^{d \times d}$ is the output projection matrix.

To enhance the stability and efficiency of training, we add a residual connection into the model architecture.

$$H'' = \text{LayerNorm}(H' + H) \quad (5)$$

3.3.4 Mean Pooled Sequence Representation. Mean pooling is done on the token embeddings in order to obtain a fixed-size representation of the sequence. The mean pooling works by averaging the values of the token embeddings, ensuring the resulting vector captures the sequence's overall characteristics. Moreover, it minimizes the effect of single tokens, making it possible for the model to pay attention to the

whole sequence. This technique proves efficient in obtaining a brief and fixed-size representation of the sequence.

$$P = \frac{1}{L} \sum_{i=1}^L h_i'' \quad (6)$$

The resulting vector $P \in \mathbb{R}^d$ serves as the global representation of the input sequence.

3.3.5 Classification Layer for Sarcasm Detection. To avoid overfitting, a dropout layer is used on the pooling vector P . Another dense layer, which is completely connected, processes the vector after dropout. This layer transforms the pooled vector into logits, which are class-specific scores. It enables the model to make predictions on all classes. To increase the robustness of prediction, we use dropout before classifying, and a fully connected layer provides raw prediction values for both sarcasm and non-sarcasm labels.

$$Z = W_F P + b, \quad Z \in \mathbb{R}^C \quad (7)$$

where $W_F \in \mathbb{R}^{d \times C}$ is the weight matrix, $b \in \mathbb{R}^C$ is the bias vector, $C = 2$ is the number of output classes (Sarcastic, Non-Sarcastic). The softmax function is used to forecast the probabilities of the classes:

$$\hat{y} = \text{softmax}(Z) \quad (8)$$

3.3.6 Training with Weighted Cross-Entropy Loss. During training, a weighted cross-entropy loss function is used to handle the imbalance of the dataset:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N w[y_i] \log(\hat{y}_{i,y_i}) \quad (9)$$

where y_i is the true label of the i -th sample, $\hat{y}_{i,y_i}$ is the predicted probability of the true class, $w[y_i]$ is the class weight for y_i , N is the batch size.

3.3.7 Optimizing with AdamW and Learning Rate Scheduling. Model parameters are optimized using AdamW optimizer, which is a variant of Adam that decouples weight decay from the d=gradient update. In addition to the optimizer, we employ a learning rate scheduler to control the pace of learning over time. Initially, a warmup phase is applied to gradually increase the learning rate, which helps stabilize early training and prevent large, unstable updates. This combination of AdamW and learning rate scheduling improves both training efficiency and final accuracy.

$$\eta_t = \eta_0 \cdot \text{cosine_annealing}(t) \quad (10)$$

where η_0 is the initial learning rate and t is the current training step.

3.3.8 Evaluating with K-Fold Cross-Validation. The K-Fold Cross-Validation technique is utilized to ensure reliability and universality because the model is trained using $K - 1$ folds and tested using the last fold. In addition, among those included in the evaluation metrics are Loss, F1-Score, and Accuracy. The mean of the parameters obtained from each fold represents the final result.

3.4 Overall algorithm of the proposed model

The algorithmic processes of the suggested sarcasm prediction model, SarcasmBERTa, are demonstrated in this sub-section. The suggested algorithm is as follows:

Step 1, The raw datasets are loaded to clean them as they consist of emojis, unnecessary spaces, paragraph breaks, etc. There are a total of two types of datasets, sarcastic and non-sarcastic. clean the raw text by removing irrelevant characters and optionally lowercasing, then concatenate two types of datasets. The RoBERTa tokenizer then converts text into subword tokens, adds special tokens like <s> and </s>, and maps them to integer IDs. Both are either padded or truncated to a fixed length L , and then an attention mask is constructed (padding = 0, actual token = 1). For example, the word “sarcasm” may be tokenized into input IDs like [45, 123, 67, ...], resulting in `input_ids` and `attention_mask` as the numerical input for the model. Step 2, after tokenizing the data, we define a custom Dataset class that loads `input_ids`, `attention_mask`, and corresponding labels. The `__getitem__` method returns a tuple (x_i, y_i) for each sample i , while `__len__` returns the total number of samples in the dataset. This dataset is then wrapped with a DataLoader, enabling shuffling and batching to generate mini-batches $\{(x_{i_j}, y_{i_j})\}_{j=1}^B$. This structure facilitates efficient parallel data loading and training during model optimization.

Step 3, the data is split into k folds, to robustly estimate model performance and reduce variance. One set is employed as validation set in each fold iteration, and the rest of the $K - 1$ folds are utilized as a set of training data. This yields a range of models and evaluations. To validate across all folds, every data point is utilized exactly once. The average (or vote/average of forecasts) across all folds may be a final performance.

Step 4, insert tokenized dataset.

Step 5: RoBERTa-base has 12 transformer layers, where $d = 768$. Processes input embeddings through self-attention and feed-forward layers, outputting contextual representations. Final [CLS] token state used for classification. It is a custom model as custom multi-head attention is added.

Steps 5.1 to 5.5, customizing the pre-trained RoBERTa model with custom MHA mechanism. The MHA mechanism computes scaled dot-product attention using query (Q), key (K), and value (V) matrices. Each head splits these matrices into smaller projections, applies attention independently, and combines the results. The attention output for each head is computed as $\text{head}_i = \text{softmax}\left(\frac{QW_i^Q(KW_i^K)^\top}{\sqrt{d_k}}\right) V W_i^V$, where d_k is the head dimension. The final multi-head output concatenates all heads and applies a linear projection. This parallel processing enables the model to capture diverse token relationships efficiently. For each epoch in each fold, the loop alternates training and

validation phases. In training mode, batches are forwarded through the model, loss is computed and backpropagated, and optimizer steps update the weights. During training, the model that performs best on validation data (e.g. highest F1 or lowest validation loss) is saved. This avoids using an overfit model from a later epoch. Instead of always using the final epoch's weights (which may overfit), we keep the "best" checkpoint. This ensures the chosen model is the one that is generalized best on validation data. If training continues past that point, weights will deviate and performance may drop. The practice of retaining the best model according to validation results is standard.

Step 6, after training the model is tested on unseen data. Compute the same metrics as on validation (Accuracy, F1, MCC) over the test labels. These final scores report the pipeline's performance on unseen data.

Step 7, The Receiver Operating Characteristic Curve for binary classification problems is defined as the True Positive Rate against the False Positive Rate for different thresholds.

Algorithm: RoBERTa + Custom Multi-Head Attention for Sarcasm Detection

```

1: Data Preprocessing and Tokenization using Equation (1)
2: Custom PyTorch Dataset and DataLoader
3: K-Fold Cross Validation Setup
4: Dataset  $D = \{(x_i, y_i)\}_{i=1}^N$ , learning rate  $\eta$ , epochs  $E$ , heads  $h$ 
5: Load the Custom RoBERTa Model and Training
for Loop over the total number of training epochs do
    for Iterate over batches of input data  $x_i$  with corresponding labels  $y_i$ . do
        Pass the batch through RoBERTa to get token embeddings of shape
         $\mathbb{R}^{b \times T \times d}$  using Equation (2).
        for For each attention head  $j$  : do
            Compute head $_j$  using scaled dot-product attention with  $Q_j, K_j$ , and
             $V_j$ 
        end
        5.1: Concatenate all attention heads' outputs using Equation (3)
        Multiplication by output weight matrix  $W^O$  to get the combined
        attention representation  $Z$ 
        5.2: Extract the representation of the [CLS] token from the combined
        output  $Z$ ,
        5.3: Apply a linear layer with weights  $W_c$  and bias  $b_c$  using Equation (8)
        5.4: Calculate weighted binary cross-entropy loss  $L$  using Equation (9)
        5.5: Calculate Validation F1 Score
        if The current validation F1 score > The previous best: then
            Update the best validation F1 score.
            save the model
        end
    end
end

```

return the trained model f_θ and the last batch predictions $\hat{y}$

6: Generate Reports

7: Analyze results based on various evaluation metrics

4 Dataset and Evaluation Metrics

There are a total of five datasets. These datasets are for English, Low Resource Language, Hindi, and Bengali. The descriptions of the datasets are as follows:

4.1 English

33,859 English tweets manually annotated as 1 for sarcasm and 0 for non-sarcasm. The data is cleaned up by removing columns, hence only retaining the required columns. There are 17,598 non-sarcastic tweets, and 16,261 tweets are sarcastic [10].

4.2 Bengali.

Lora et al. [34] create a dataset 'Ben-Sarc' (A Dataset for Sarcasm Detection), 25,636 balanced datasets combination of equal division of sarcastic and non-sarcastic datasets.

4.3 Hindi

Katyayan introduced [27] a collection of a total of 16,000 Hindi tweets, where 6,051 tweets are sarcastic and 10,128 tweets are non-sarcastic. Kumar et al. include [30] "Panchayat" – a Hindi web series from HumourHindiNet, which offers 3,000+ humor-rich dialogues. "Kota Factory" – another Hindi web series from HumourHindiNet – consists of 2,000+ sarcastic dialogues.

4.4 Evaluation Metrics

In the current section, we provide justification for various metrics used for their performance measures based on our recommended algorithm. The explanations for these metrics are mentioned below:

- i. **Accuracy (Acc):** Accuracy reflects the ratio of correctly predicted instances out of the total. In simple terms, it tells you how often the predictions are right. Mathematically, it is:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (11)$$

- ii. **Precision (Pre):** Precision tells us how many of the things we call positive are actually correct. In simple terms, it shows the accuracy of our positive predictions. Mathematically, we write it as:

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

- iii. **Recall (R):** Recall is a performance metric that calculates the ratio of genuinely positive instances labeled correctly. Recall can be formulated mathematically for a binary classification problem as:

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

- iv. **F1 Score (F1):** Harmonic mean (F1 Score) refers to the compromise between recall and precision. This means that the performance of the model is determined by taking both accuracy and ability to capture all positive examples into consideration. The formula used to determine it is:

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (14)$$

- v. **Consistency Loss (CL):** Consistency Loss calculates the reliability of predictions made by the model in case of minor changes in the input text. For sarcasm detection models, minor changes in word choice and representations should not affect the output class predictions, provided that the semantics remain unchanged. Thus, consistency loss is the measure of dissimilarity between the probability distributions predicted by the model on the input and its perturbation. Let x be the input text, and let $\tilde{x}$ denote a perturbation of the same text obtained using the method of random dropout through forward propagation. The probability distributions corresponding to the respective inputs are denoted as $p(x)$ and $p(\tilde{x})$. This loss is shown below:

$$CL = ||p(x) - p(\tilde{x})||^2 \quad (15)$$

When the consistency loss takes a smaller number, this means that the model provides consistent predictions against small disturbances, which suggests that it is more robust.

- vi. **MCC:** It ranges from -1 for inverse prediction to $+1$ for perfect prediction in the case of binary classification. A perfect prediction is shown by $+1$, an average random guess is shown by 0 , and an inverse prediction is shown by -1 . Matthews Correlation Coefficient (MCC) is particularly useful in cases of unbalanced classes because it considers the imbalance in the calculation. Its mathematical expression is:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

- vii. **ROC Curve (Receiver Operating Characteristic Curve):** is a graphical representation of the performance of a binary classification model for all classification thresholds. Using the ROC curve, we can now also compare different classification methods. Mathematically it is written as:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \quad (17)$$

The more vertical the line is, the better the categorization model is. The more accurate the classifier is, the larger the area under the curve is. The numerical value of the AUC represents the area under the curve.

Here, **True Positive (TP)** means an instance is correctly identified. **True Negative (TN)** means an instance is correctly rejected. **False Positive (FP)** happens when something is incorrectly identified, and **False Negative (FN)** happens when something is incorrectly rejected.

5 Experimental Results

Here, we explained the experimental comparison of the methodology developed. We performed experiments on the data sets given in Section 4, and performance was measured by metrics as mentioned in Section 5. This enabled us to evaluate how well our proposed method identifies negative stances in multilingual social media posts. In addition, we compared our approach with several baseline algorithms, including TextCNN [13], and Bi-LSTM [18], to gain a more comparative study.

Evaluation of the model was performed against a variety of existing modern-day sarcasm detection models, including those proposed by Zhang et al. [51], Tay et al. [50], Devlin et al. [28], Xiong et al. [51], and Nguyen et al. [38]. Table 1 illustrates the hyperparameters used to enhance the performance of the model. The batch size used in this case was set to 16 for efficiency, and the max length was set to 256 to ensure context. The learning rate was set to 1e-5, and the model was trained for 50 epochs. For efficient training, 100 warm-up steps were used. The dropout rate was set to 0.3 to ensure that overfitting was avoided. For the multi-head attention, the number of heads was set to eight. The model used the RoBERTa model to perform the role of a transformer. The model also used multi-head attention to ensure that it effectively extracted all the information from the text. In the context of effective training process stability, the model utilizes the concept of ‘Layer Normalization.’ For the purpose of efficient regularization, the model utilizes regularization. In the context of the optimization process for the fine-tuning phase, the model utilizes the concept of ‘AdamW.’ A ‘cosine annealing’ learning rate scheduler is utilized to efficiently plan the learning rate. In the context of effectively addressing the ‘class imbalance,’ the model utilizes the concept of ‘class weighting.’

Table 1. Hyperparameters for SarcasmBERTa on Datasets 1 to 5

Hyperparameter	English (D1)	Bengali (D2)	Hindi (D3)	Hindi (D4)	Hindi (D5)
Max Length	256	256	256	256	256
Batch Size	16	16	16	16	16
Epochs	50	100	30	50	50
Warmup Steps	100	100	100	100	100
Learning Rate	1e-5	1e-5	1e-5	1e-5	1e-5
Dropout Rate	0.3	0.3	0.3	0.3	0.3
Number of Heads	8	8	8	8	8
Class Weights	[0.7, 1.3]	[0.7, 1.3]	[0.7, 1.3]	[0.7, 1.3]	[0.7, 1.3]
Number of Folds (k)	5	5	5	5	5

5.1 Hyperparameter Study

A specific number of parameters is used, which include batch size, epochs, warmup steps, learning rate, etc. In this section, the hyperparameters of the model are examined. Max length represents the input token’s maximum sequence length. When this value is increased, Batch size is needed when max length becomes high. When the batch size is high, it enables the model to learn better without using up a lot of memory. The number

of times the model iterates over the entire dataset is determined by epochs, which helps in preventing overfitting using early stopping or validation monitoring. The gradual increase in the learning rate using the number of steps in warmup is used to prevent sudden changes in the model weights. When a lower learning rate is chosen, Fine-tuning of pre-trained models like BERT or RoBERTa, which is beneficial in retaining the knowledge that is already learned in the model. Dropout is a regularization method where some of the input elements are randomly set to zero during training. This is beneficial in preventing overfitting by not letting the model get too dependent on certain neurons or features. Representational ability is improved in the model through the use of attention heads, enabling it to pay attention to various parts of the input at once. Class imbalance is addressed by assigning weights to each class within the loss function in order to improve classification performance. AdamW is an optimizer variant that stands for Adam with Weight Decay. It is beneficial in dealing with L2 regularization or weight decay, especially in transformer-based models like BERT, RoBERTa, etc. The scheduler is used to control the learning rate during training, which is beneficial in helping the model converge faster with high accuracy.

5.2 Performance comparison with Baseline Algorithms and Contemporary Models

This section sought to assess how well the proposed algorithm performed in comparison to a number of baseline algorithms, including TextCNN, Bi-LSTM, and GloVe. To achieve this, the baseline algorithms are refined for optimal performance. The performance of the proposed algorithm was compared to a number of existing algorithms as part of the experimental study, such as those by P. Zhang et al.[51], Y. Tay et al. [50], J. Devlin et al. [28], and D. Q. Nguyen et al. [38]. Table 3 shows the Correct Rate, Exactness, Sensitivity results, and Harmonic Mean respectively for these algorithms and ours over the entire set of datasets 1. Analysis of these results demonstrates that our algorithm ranks highest in terms of performance across all datasets and metrics. In Dataset 1, for the English datasets we used Matthews' correlation coefficient (MCC), for each test fold. The model achieved consistently high MCC scores. The highest MCC score was 0.9211, reflecting the model accurately identified sarcastic and non-sarcastic cases. In order to do the additional analysis of the consistency of the SarcasmBERTa model, Consistency Loss (CL) will be the new performance metric that will be taken into account along with Accuracy, Precision, Recall, and F1-score. Consistency Loss is calculated according to how varied the probabilities of the original text and its altered form are when making a prediction. Given the importance of the context in understanding sarcasm, a good sarcasm detection model must have consistent predictions even after minor changes made to the original input data. This means that smaller CL values will show more consistent predictions. Furthermore, to compare the model with multimodal and hybrid approaches presented recently, MMFN [39] and KFGC-Net [58] were evaluated.

The performance metrics of Accuracy, Precision, Recall, and F1-score used by the proposed SarcasmBERTa model and the baseline algorithms used in the experiments

on real-world datasets are presented in Table 2. In the Table 2, Table 3, Table 4, and Table 5,

Table 2. Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Dataset 1 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).

Dataset	Method	Acc	Pre	R	F1	CL
English	TextCNN (Chen, 2015)	0.80	0.74	0.76	0.75	0.052
	SIARN (Tay et al., 2018)	0.80	0.75	0.75	0.76	0.049
	SMSD (Xiong et al., 2019)	0.81	0.76	0.75	0.76	0.047
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.82	0.77	0.78	0.77	0.046
	GloVe (Pennington et al., 2014)	0.82	0.81	0.82	0.81	0.045
	BERT (Devlin et al., 2019)	0.84	0.79	0.82	0.80	0.041
	BERTweet (Nguyen et al., 2020)	0.88	0.87	0.88	0.88	0.034
	KnowleNet (Yue et al., 2023)	0.89	0.88	0.84	0.86	0.029
	KFGC-Net (Zhuang et al., 2025)	0.72	0.66	0.70	0.68	0.059
	MMFN (Ou and Li, 2025)	0.78	0.78	0.76	0.77	0.049
	SarcasmBERTa	0.97	0.96	0.96	0.96	0.019

5.3 Evaluation on Bengali Datasets

To evaluate our model's performance in Bangla datasets, SarcasmBERTa was used. The capability and scalability of Sarcasm detection in a unique language and culture are provided. After performing the experiments on our model, it provided high accuracy and precision rates. The experiments performed on our model showed its robustness. Most importantly, it was capable of extracting Sarcasm cues from English and Bangla texts effectively, overcoming the hurdles of different language structures and culture. These experiments and results validate our approach and highlight the scalability of our approach for multilingual Sarcasm classification. The model performed greatly, achieving remarkable results despite syntactic and semantic differences from English. These results validate our approach and highlight its scalability for multilingual Sarcasm detection. In this paper, Indic-Transformers Bengali BERT [24] is also referred to as a short form of 'I-T Bengali BERT'. Table 3 showed the comparison with other models and our proposed model.

5.4 Evaluation on Hindi Datasets

SarcasmBERTa's ability to detect sarcasm across different languages was examined using the Hindi Tweets Dataset-3 for Sarcasm Detection, which contained over 16,000 tweets. The dataset [27] included 6,051 sarcastic tweets and 10,128 non-sarcastic tweets, making it a rich resource for sarcasm detection in Hindi. We used this dataset to test SarcasmBERTa's ability to generalize across the linguistic and cultural aspects of Hindi. The efficiency of SarcasmBERTa in recognizing sarcasm in various conditions was evident from its high accuracy and precision rate. The effectiveness of adapting

Table 3. Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Dataset 2 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).

Dataset	Method	Acc	Pre	R	F1	CL
Bengali	TextCNN (Chen, 2015)	0.65	0.64	0.65	0.64	0.061
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.66	0.65	0.66	0.65	0.059
	GloVe (Pennington et al., 2014)	0.66	0.66	0.66	0.66	0.058
	BERT (Devlin et al., 2019)	0.68	0.69	0.64	0.66	0.054
	I-T Bengali BERT (Jain et al., 2020)	0.75	0.74	0.77	0.75	0.041
	KFGC-Net (Zhuang et al., 2025)	0.70	0.68	0.69	0.68	0.058
	MMFN (Ou and Li, 2025)	0.73	0.72	0.72	0.72	0.052
	SarcasmBERTa	0.82	0.82	0.82	0.82	0.026

the model to the unique syntax and semantics of the Hindi language was evidenced through the results, where the effectiveness of using the approach for detecting sarcasm across multiple languages was clearly demonstrated. It was observed that SarcasmBERTa performed excellently on the dataset. This is came to notice that, SarcasmBERTa performed exceptionally well on the data set, and the results were great. The results proved that this approach was sufficiently powerful to detect sarcasm in various complex contexts. In addition, the results proved that the model had the potential to adapt well to the specific syntactic and semantic characteristics of the Hindi language. The results shown in Table 5 proved that the model was generalizable and successful in multilingual sarcasm detection. We used two additional datasets: Dataset 4 to further evaluate the performance of SarcasmBERTa, which was based on the Hindi web series *Panchayat*, and Dataset 5, based on *Kota Factory*. The data sets used were for the purpose of humor detection, with utterances being classified as either ‘Humor’ or ‘Non-Humor.’ The data sets selected were appropriate because they relate to sarcasm and humor, allowing us to assess SarcasmBERTa’s ability to differentiate between them. The robustness of the model in identifying sarcasms was affirmed upon scrutinizing the data sets, which indicated that the model had excellent accuracy and precision in identifying them in both cases. The comparison of the performance of our proposed model, Sarcasm-BERTa, and other models, both old and new, was demonstrated in Table 4. SarcasmBERTa demonstrated its efficiency in handling all datasets, including English, Bengali, and Hindi.

The evaluation parameters to measure the efficacy of our model, as shown in Table 5, were:

Figure 3 shows loss curves, that there is a general trend of decreasing loss throughout the 50 epochs. Within the first 10 epochs, the loss decreased rapidly from around 0.5 to less than 0.2. This shows that the network is learning and adapting very well. At the same time, the validation loss is also decreasing, albeit with some fluctuations. In the second part of the graph, between epochs 11 and 30, the network continues to reduce its loss steadily, now to below 0.1. At the same time, the validation loss is

Table 4. Performance comparison of the proposed SarcasmBERTa model with baseline and contemporary models on Datasets 3, 4, and 5 using Accuracy (Acc), Precision (Pre), Recall (R), F1-score (F1), and Consistency Loss (CL).

Dataset	Method	Acc	Pre	R	F1	CL
Hindi	TextCNN (Chen, 2015)	0.79	0.73	0.75	0.74	0.053
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.80	0.77	0.78	0.77	0.051
	GloVe (Pennington et al., 2014)	0.81	0.78	0.79	0.78	0.048
	BERT (Devlin et al., 2019)	0.82	0.80	0.81	0.80	0.044
	HumourHindiNet (Kumar et al., 2024)	0.84	0.83	0.84	0.83	0.039
	KFGC-Net (Zhuang et al., 2025)	0.78	0.76	0.77	0.76	0.050
	MMFN (Ou and Li, 2025)	0.81	0.79	0.80	0.79	0.044
	SarcasmBERTa	0.96	0.96	0.96	0.96	0.021
Kota Factory	TextCNN (Chen, 2015)	0.78	0.74	0.76	0.75	0.056
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.82	0.77	0.78	0.77	0.052
	BERT (Devlin et al., 2019)	0.89	0.93	0.86	0.89	0.038
	GloVe (Pennington et al., 2014)	0.90	0.95	0.83	0.89	0.036
	HumourHindiNet (Kumar et al., 2024)	0.91	0.96	0.87	0.91	0.033
	KFGC-Net (Zhuang et al., 2025)	0.82	0.79	0.80	0.79	0.047
	MMFN (Ou and Li, 2025)	0.85	0.82	0.83	0.82	0.041
	SarcasmBERTa	0.96	0.92	0.91	0.92	0.022
Panchayat	TextCNN (Chen, 2015)	0.77	0.75	0.77	0.76	0.058
	Bi-LSTM (Graves and Schmidhuber, 2005)	0.80	0.78	0.79	0.78	0.053
	GloVe (Pennington et al., 2014)	0.85	0.83	0.86	0.85	0.046
	BERT (Devlin et al., 2019)	0.86	0.84	0.87	0.86	0.041
	HumourHindiNet (Kumar et al., 2024)	0.87	0.90	0.84	0.87	0.038
	KFGC-Net (Zhuang et al., 2025)	0.84	0.81	0.82	0.81	0.046
	MMFN (Ou and Li, 2025)	0.87	0.84	0.85	0.84	0.039
	SarcasmBERTa	0.94	0.89	0.91	0.90	0.024

steady between 0.2 and 0.3. This shows that the network is becoming more consistent in its performance. Finally, in the third part of the graph, between epochs 31 and 50, the loss is very low, now below 0.05, and the validation loss is also low at around 0.2. This shows that the network has been able to minimize its errors without the risk of overfitting.

In Figure 4, the training and validation accuracy across 50 epochs showed consistent improvement. During the first phase, from epochs 1 to 10, the training accuracy rose from around 73% to over 90%, demonstrating the model's ability to learn quickly. At the same time, validation accuracy increased from 78% to about 88%, with only minor fluctuations. In the middle phase, from epochs 11 to 30, the training accuracy exceeded 96%, while the validation accuracy grew steadily, crossing 90%. The gap between the two metrics was very small, indicating that the model generalized well. In the final phase, from epochs 31 to 50, the training accuracy reached nearly 98%, and the validation accuracy reached 96.10% at the end of the 50 epochs. The convergence of both metrics confirmed improvement of our model and the absence of overfitting. The

Table 5. A Study of SarcasmBERTa on datasets and Evaluation metrics

Dataset Type	Description	Evaluation Metric Calculated
English (Dataset 1)	English tweets (33,859): 17,598 non-sarcastic, 16,261 sarcastic.	Accuracy: 0.96, Precision: 0.96, Recall: 0.96, F1 Score: 0.96, Consistency loss: 0.019
Bengali (Dataset 2)	Ben-Sarc, containing 25,636 comments, 12,818 non-sarcastic, 12,818 sarcastic	Accuracy: 0.82, Precision: 0.825, Recall: 0.82, F1 Score: 0.82, Consistency loss: 0.026
Hindi (Dataset 3)	Hindi tweets consist of 16,000 tweets where 6,051 sarcastic tweets and 10,128 non-sarcastic.	Accuracy: 0.97, Precision: 0.95, Recall: 0.95, F1 Score: 0.95, Consistency loss: 0.021
Hindi (Dataset 4)	Hindi web series Panchayat 3,000 plus datasets.	Accuracy: 0.96, Precision: 0.92, Recall: 0.91, F1 Score: 0.92, Consistency loss: 0.022
Hindi (Dataset 5)	Hindi web series Kota-Factory 2,000 plus datasets.	Accuracy: 0.94, Precision: 0.89, Recall: 0.91, F1 Score: 0.90, Consistency loss: 0.024



Fig. 3. Representation of Training and Validation Loss over Epochs

confusion matrix (Fig. 5) showed high level of performance of our proposed model. Both the TP and TN values were well above 90%. Specifically, it was observed that the model correctly classified 95.79% of sarcastic instances and 96.41% of non-sarcastic ones. The FP and FN rates were also very low, at 3.48% and 4.84%, respectively. Misclassifications rates between sarcastic and non-sarcastic texts were minimal, reflecting its competence in recognizing nuanced language patterns. A ROC curve was created to check the performance of the sarcasm detection model. It showed a strong separation between the true positive rate (TPR) and the false positive rate (FPR), indicating good performance and excellent classification capability. The area under the curve (AUC) is 0.9870.

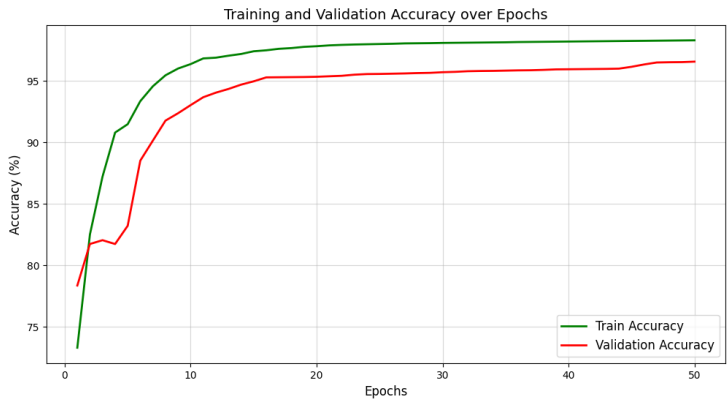


Fig. 4. Representation of Training and Validation Accuracy over Epochs

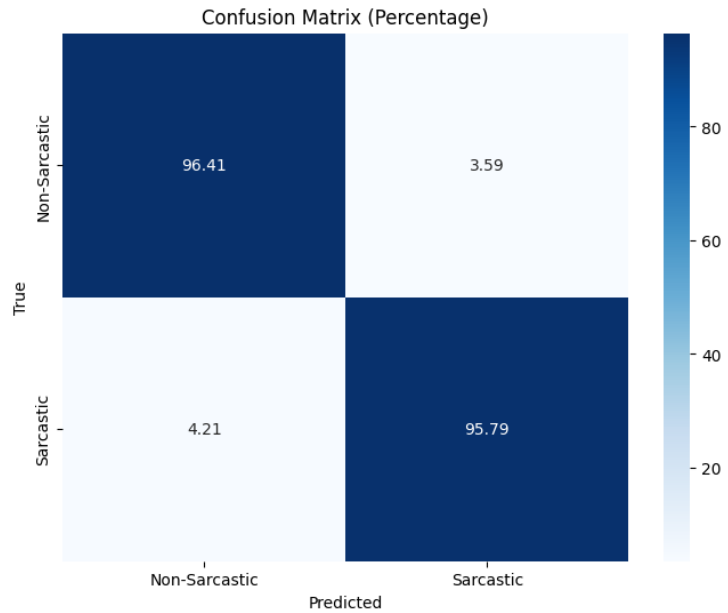


Fig. 5. Confusion Matrix for Sarcasm Detection Model

5.5 Ablation Study

An ablation research was carried out with the aim of focusing on the effect of the suggested custom multi-head attention mechanism in order to effectively evaluate the contribution of the suggested components. For this purpose, different improved versions with the incorporation of attention processes were compared with the suggested

fine-tuning of the RoBERTa method. The aim behind this research is to evaluate the effect of the suggested custom multi-head attention layer on the model’s ability to effectively recognize contextual sarcastic signals in multilingual datasets. In addition, a consistency regularization term was incorporated during training to encourage stable predictions for perturbed inputs. This regularization further improved the robustness and generalization capability of the proposed SarcasmBERTa (RoBERTa + Custom Multi-Head Attention) model. Table 6 shows the ablation research results.

Table 6. An ablation study to evaluate the impact of the attention mechanism used in the SarcasmBERTa model.

Model Variant	Correct Rate	Harmonic Mean
RoBERTa (Baseline Fine-tuning)	0.92	0.92
RoBERTa + Attention Layer	0.94	0.94
SarcasmBERTa	0.96	0.96

6 Conclusion

The misinterpretation of sarcastic content on social networks can lead to the failure of sentiment analysis, which can hinder the achievement of effective human-computer interactions in online communication, such as social networking sites including Facebook and Twitter. The ability to identify sarcasm helped to achieve effective communication in multilingual environments. Therefore, the aim of this research was to propose a novel approach to the detection of sarcasm in multilingual environments using the MHA mechanism and the advanced RoBERTa model. Moreover, the reliability and robustness of the proposed model, referred to as SarcasmBERTa, were ensured using the k-fold cross-validation method. The results proved the effectiveness of our approach in detecting sarcasm in the context of the experiment. The impressive average accuracy of our approach was 96.10%, and the F1 score was 96.15%. The results proved that the proposed SarcasmBERTa approach was successful in recognizing the contextual patterns of sarcasm in various languages. The SarcasmBERTa approach is beneficial for improving the overall performance of the sarcasm detection approach. The SarcasmBERTa approach is beneficial because it incorporates the strengths of the RoBERTa approach and the complex multi-head attention mechanism. The SarcasmBERTa approach also benefits from the robust training objectives. The SarcasmBERTa model has provided new research avenues in the development of effective NLP models that are beneficial in detecting sarcasm. The SarcasmBERTa approach can be taken further in the context of integrating the strengths of the strong language models. The SarcasmBERTa approach can be taken further in the context of developing multimodal-based sarcasm detection models that incorporate the strengths of the image and text modalities. The SarcasmBERTa approach can be taken further in the context of integrating the strengths of the cross-lingual transfer learning approach. The SarcasmBERTa approach can be

taken further in the context of developing effective NLP models that are beneficial in detecting the difficulty in sarcasm detection. SarcasmBERTa can be taken further in the context of integrating the strengths of the strong training objectives such as the consistency loss function.

References

- [1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*. 30–38.
- [2] Iram Ali Ahmad, Praveen Gatla, and Rajesh Kumar Mundotiya. 2025. Sarcasm identification and classification in Hindi newspaper headlines. *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, 4 (2025), 1–21.
- [3] Ravinder Ahuja and Subhash Chander Sharma. 2022. Transformer-based word embedding with CNN model to detect sarcasm and irony. *Arabian Journal for Science and Engineering* 47, 8 (2022), 9379–9392.
- [4] Ramisa Anan, Tasnim Sakib Apon, Zeba Tahsin Hossain, Elizabeth Antora Modhu, Sudipta Mondal, and MD Golam Rabiul Alam. 2023. Interpretable bangla sarcasm detection using bert and explainable ai. In *2023 IEEE 13th annual computing and communication workshop and conference (CCWC)*. IEEE, 1272–1278.
- [5] Girma Yohannis Bade, Jose Luis Oropeza, and Olga Kolesnikova. 2025. Pragmatic Generalization in LLMs: Insights from Fine-Tuning and Evaluating on Multilingual Sarcasm. In *Mexican International Conference on Artificial Intelligence*. Springer, 218–230.
- [6] Yongtang Bao, Xin Zhao, Peng Zhang, Yue Qi, and Haojie Li. 2025. HIAN: A hybrid interactive attention network for multimodal sarcasm detection. *Pattern Recognition* 164 (2025), 111535.
- [7] Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1363–1375.
- [8] Kheira Zineb Bousmaha, Khaoula Hamadouche, Hadjer Djouabi, and Lamia Hadrach-Belguith. 2024. Automatic algerian sarcasm detection from texts and images. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 7 (2024), 1–25.
- [9] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2506–2515. doi:10.18653/v1/P19-1239
- [10] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2506–2515.
- [11] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).
- [12] Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhat-tacharyya. 2022. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems* 257 (2022), 109924.
- [13] Yahui Chen. 2015. *Convolutional neural network for sentence classification*. Master’s thesis. University of Waterloo.
- [14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Online, 8440–8451. <https://aclanthology.org/2020.acl-main.747>
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*,

- volume 1 (long and short papers). 4171–4186.
- [16] Lori J Ducharme. 1994. Sarcasm and interactional politics. *Symbolic interaction* 17, 1 (1994), 51–62.
 - [17] Marta Dynel. 2009. Beyond a joke: Types of conversational humour. *Language and linguistics compass* 3, 5 (2009), 1284–1299.
 - [18] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610. doi:10.1016/j.neunet.2005.06.042 IJCNN 2005.
 - [19] M Greeshma and Philomina Simon. 2024. Bidirectional Gated Recurrent Unit with Glove Embedding and Attention Mechanism for Movie Review Classification. *Procedia Computer Science* 233 (2024), 528–536.
 - [20] Hunter Gregory, Steven Li, Pouya Mohammadi, Natalie Tarn, Rachel Draelos, and Cynthia Rudin. 2020. A transformer approach to contextual sarcasm detection in twitter. In *Proceedings of the second workshop on figurative language processing*. 270–275.
 - [21] Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun, Hongyi Huang, Guanghao Chen, Yitong Xu, and Yang Yang. 2022. X-pudu at semeval-2022 task 6: Multilingual learning for english and arabic sarcasm detection. *arXiv preprint arXiv:2211.16883* (2022).
 - [22] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413* (2018).
 - [23] Nivin A Helal, Ahmed Hassan, Nagwa L Badr, and Yasmine M Afify. 2024. A contextual-based approach for sarcasm detection. *Scientific Reports* 14, 1 (2024), 15415.
 - [24] Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages. *arXiv preprint arXiv:2011.02323* (2020).
 - [25] Tanya Jain, Nilesh Agrawal, Garima Goyal, and Niyati Aggrawal. 2017. Sarcasm detection of tweets: A comparative study. In *2017 Tenth International Conference on Contemporary Computing (IC3)*. IEEE, 1–6.
 - [26] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 1–22.
 - [27] Pragya Katyayan. 2020. Tweets Dataset for Sarcasm Detection in Hindi. <https://github.com/>. GitHub repository.
 - [28] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. Minneapolis, Minnesota.
 - [29] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579* (2017).
 - [30] Akshi Kumar, Abhishek Mallik, and Sanjay Kumar. 2024. HumourHindiNet: Humour detection in Hindi web series using word embedding and convolutional neural network. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 7 (2024), 1–21.
 - [31] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
 - [32] Bin Liang, Lin Gui, Yulan He, Erik Cambria, and Ruifeng Xu. 2024. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing* (2024).
 - [33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
 - [34] Sanzana Karim Lora, G. M. Shahariar, Tamanna Nazmin, Noor Nafeur Rahman, Rafsan Rahman, Miyad Bhuiyan, and Faisal Muhammad Shah. 2025. Ben-Sarc: A self-annotated corpus for sarcasm detection from Bengali social media comments and its baseline evaluation. *Natural Language Processing* 31, 2 (2025), 674–699. doi:10.1017/nlp.2024.11

- [35] Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems* 34, 3 (2019), 38–43.
- [36] Jiana Meng, Yanlin Zhu, Shichang Sun, and Dandan Zhao. 2024. Sarcasm detection based on BERT and attention mechanism. *Multimedia Tools and Applications* 83, 10 (2024), 29159–29178.
- [37] Abhijit Mishra and Pushpak Bhattacharyya. 2016. Deep Learning for Sarcasm Detection in Tweets. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California, 100–105.
- [38] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200* (2020).
- [39] Lisong Ou and Zhixin Li. 2025. Multi-modal sarcasm detection on social media via multi-granularity information fusion. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 3 (2025), 1–23.
- [40] Rajnish Pandey and Jyoti Prakash Singh. 2023. BERT-LSTM model for sarcasm detection in code-mixed social media post. *Journal of Intelligent Information Systems* 60, 1 (2023), 235–254.
- [41] Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eysers. 2021. Detecting the target of sarcasm is hard: Really?? *Information processing & management* 58, 4 (2021), 102599.
- [42] Himani Pokhriyal and Goonjan Jain. 2025. Supposititious sarcasm detection and sentiment analysis coping hindi language in social networks harnessing zipf-mandelbrot probabilistic optimisation and perplexity entropy learning. *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, 2 (2025), 1–28.
- [43] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. The COLING 2016 Organizing Committee, Osaka, Japan, 1601–1612. <https://aclanthology.org/C16-1151/>
- [44] Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3930–3940.
- [45] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2. 0: Towards a reliable multi-modal sarcasm detection system. *arXiv preprint arXiv:2307.07135* (2023).
- [46] Dania Refai, Maged S. Al-Shaibani, and Irfan Ahmad. 2025. Is This the Best Prompt? Scoring Prompts for Arabic NLP Across LLMs. *IEEE Access* 13 (2025), 171468–171492. doi:10.1109/ACCESS.2025.3616181
- [47] Axel Rodríguez, Yi-Ling Chen, and Carlos Argueta. 2024. SINCERE: A hybrid framework with graph-based compact textual models using emotion classification and sentiment analysis for Twitter sarcasm detection. *IEEE Transactions on Computational Social Systems* 11, 5 (2024), 5593–5606.
- [48] Mochamad Alfan Rosid, Daniel Siahaan, and Ahmad Saikhu. 2024. Sarcasm Detection in Indonesian-English Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi-Directional GRU. *IEEE Access* (2024).
- [49] Dilip Kumar Sharma, Bhuvanesh Singh, Saurabh Agarwal, Hyunsung Kim, and Raj Sharma. 2022. Sarcasm detection over social media platforms using hybrid auto-encoder-based model. *Electronics* 11, 18 (2022), 2844.
- [50] Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. *arXiv preprint arXiv:1805.02856* (2018).
- [51] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference*. 2115–2124.
- [52] 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).
- [53] Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. 2023. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion* 100 (2023), 101921.

- [54] Xiaoqiang Zhang, Ying Chen, and Guangyuan Li. 2021. Multi-modal sarcasm detection based on contrastive attention mechanism. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I* 10. Springer, 822–833.
- [55] Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. 2024. Learning Multitask Commonness and Uniqueness for Multimodal Sarcasm Detection and Sentiment Analysis in Conversation. *IEEE Transactions on Artificial Intelligence* 5, 3 (2024), 1349–1361. doi:10.1109/TAI.2023.3298328
- [56] Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025. SarcasmBench: Towards Evaluating Large Language Models on Sarcasm Understanding. *IEEE Transactions on Affective Computing* 16, 4 (2025), 2560–2578. doi:10.1109/TAFFC.2025.3604806
- [57] Juliann Zhou. 2023. An evaluation of state-of-the-art large language models for sarcasm detection. *arXiv preprint arXiv:2312.03706* (2023).
- [58] Xingjie Zhuang, Fengling Zhou, and Zhixin Li. 2025. Multi-modal sarcasm detection via knowledge-aware focused graph convolutional networks. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 5 (2025), 1–22.

Proof of Publishing

1. Proofs for the paper: Suman Saha and Sanjay Kumar, “Explainable Multi-Class Mental Health Classification from Social Media Using DistilBERT and Knowledge Graphs,” 2026 Fourth International Conference on Secure Cyber Computing and Communication (ICSCCC 2026) to be held on May 29 - 31, 2026 at NIT Jalandhar.

481 - Registration Reminder for ICSCCC-2026 Inbox x



Microsoft CMT <noreply@msr-cmt.org>
to me ▾

Apr 20, 2026, 9:54 AM ★ 😊 ↩ ⋮

Dear Dr. Suman Saha,

Greetings!

Congratulations once again on the acceptance of your paper titled Explainable Multi-Class Mental Health Classification from Social Media Using DistilBERT and Knowledge Graphs for ICSCCC-2026, scheduled to be held from 29-31 May 2026 at Dr. B. R. Ambedkar National Institute of Technology, Jalandhar.

This is a gentle reminder that at least one author must complete the registration process at the earliest to ensure inclusion of your paper in the final conference proceedings and its submission for publication to IEEE.

Please note that papers without a completed registration by the deadline will not be included in the official proceedings. As today is the last date for registration, we strongly urge you to complete the process without delay to avoid any last-minute issues.

*****If you have already registered, kindly disregard this message.

We look forward to your participation and to making the conference a great success.
Warm regards,
Team ICSCCC-2026

Please do not reply to this email as it was generated from an email account that is not monitored.

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our [Privacy Statement](#).

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052



Dr B R Ambedkar National Institute of Technology
Jalandhar - 144008, Punjab, India

**4th International Conference on
Secure Cyber Computing and communications
(May 29 - 31, 2026)**

Registration Fee e-Receipt

Participant's Details:

Registration ID	2025481158
Name	SUMAN SAHA
Category	External (Outside NITJ)
Participant Type	Student
Paper ID	481
Membership Type	Non-IEEE Member
Nationality	India
Mobile No.	8017256146
Email ID	sumansaha0178@gmail.com

Fee Details:

Amount (GST 18%)	INR 8850.00
Transaction ID	114392127336
Transaction Date	26-Mar-2026 20:13:52
Payment Status	SUCCESS

2. Proofs for the paper: Suman Saha and Sanjay Kumar, “ SarcasmBERTa: Multilingual Sarcasm Detection using RoBERTa and Multi-Head Attention Based Model,” accepted with minor revisions to ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).

← [Icons] 2 of 19 < >

TALLIP - Decision on Manuscript ID TALLIP-25-0376 Inbox x [Icons]

 **Transactions on Asian and Low-Resource Language Information Processing** <onbehalfof@manuscriptcentral.com>
to sanjay.kumar, yuvalmarton, me ▾ Mar 7, 2026, 10:26 AM ★ 😊 ↶ ⋮

06-Mar-2026

Dear Dr Kumar:

Manuscript ID TALLIP-25-0376 entitled "SarcasmBERTa: Multilingual Sarcasm Detection Using RoBERTa and Multi-Head Attention Based Model", which you submitted to the Transactions on Asian and Low-Resource Language Information Processing, has been reviewed. The comments of the reviewer(s) are included at the bottom of this letter.

The reviewer(s) have recommended publication, but also suggested some minor revisions to your manuscript. As such, your manuscript is conditionally accepted, provided that you make the revisions as recommended by the associate editor and the reviewers. Therefore, I invite you to revise your manuscript to address their comments.

Please submit your revised manuscript within one month (i.e., by 06-Apr-2026). Please follow the formatting instructions at:

<http://www.acm.org/publications/submissions>

In your revised manuscript, please append your response to the reviewers' comments at the end of your revised manuscript, and submit in one single PDF file.

To update your manuscript, log into <https://mc.manuscriptcentral.com/tallip> and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

Thank you for submitting your manuscript to the Transactions on Asian and Low-Resource Language Information Processing and I look forward to receiving your revision by the given deadline.

Sincerely,
Dr Imed Zitouni
Editor in Chief, Transactions on Asian and Low-Resource Language Information Processing

← [Icons] 1 of 19 < >

Transactions on Asian and Low-Resource Language Information Processing - Manuscript ID TALLIP-25-0376.R1 Inbox x [Icons]

 **Transactions on Asian and Low-Resource Language Information Processing** <onbehalfof@manuscriptcentral.com>
to sanjay.kumar, me ▾ Mon, Apr 20, 10:18 PM ★ 😊 ↶ ⋮

20-Apr-2026

Dear Dr Kumar:

Your manuscript entitled "SarcasmBERTa: Multilingual Sarcasm Detection Using RoBERTa and Multi-Head Attention Based Model" has been successfully submitted online and is presently being given full consideration for publication in the Transactions on Asian and Low-Resource Language Information Processing.

Your manuscript ID is TALLIP-25-0376.R1.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to Manuscript Central at <https://mc.manuscriptcentral.com/tallip> and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Center after logging in to <https://mc.manuscriptcentral.com/tallip>.

Thank you for submitting your manuscript to the Transactions on Asian and Low-Resource Language Information Processing.

Sincerely,
Transactions on Asian and Low-Resource Language Information Processing Editorial Office

ORCID

Starting in 2021, **ACM** requires that all authors provide their ORCID at acceptance (and preferably at submission). To ensure a seamless process, we suggest that you ensure that your co-authors have signed up for an ORCID as soon as possible. Registering is easy and takes less than five minutes. To register, visit Register for ORCID.

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized.

S S

Report



Quick Submit



Quick Submit



Delhi Technological University

Document Details

Submission ID

trn:oid:::1:3596305963

Submission Date

Jun 17, 2026, 1:13 PM GMT+5:30

Download Date

Jun 17, 2026, 1:23 PM GMT+5:30

File Name

Only_for_plag_check_thesis.pdf

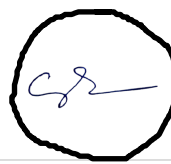
File Size

2.4 MB

47 Pages

11,351 Words

63,812 Characters



*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Report

by S S

Submission date: 17-Jun-2026 01:14PM (UTC+0530)

Submission ID: 2984848602

File name: Only_for_plag_check_thesis.pdf (2.41M)

Word count: 11351

Character count: 63812

Report



ORIGINALITY REPORT

9%

SIMILARITY INDEX

4%

INTERNET SOURCES

10%

PUBLICATIONS

1%

STUDENT PAPERS

PRIMARY SOURCES

1	Nirmal Bhandari, Sujing Wang. "Social Media Analysis for Mental Health using Large Language Models", 2025 10th International Conference on Big Data Analytics (ICBDA), 2025 Publication	1%
2	"ECAI 2020", IOS Press, 2020 Publication	<1%
3	aclanthology.org Internet Source	<1%
4	link.springer.com Internet Source	<1%
5	repository.telkomuniversity.ac.id Internet Source	<1%
6	www.ncbi.nlm.nih.gov Internet Source	<1%
7	arxiv.org Internet Source	<1%
8	Bengong Yu, Chenyue Li, Zhonghao Xi, Haoyu Wang, Shuping Zhao. "Cross-modal Interaction and Multi-view Misalignment Network for multimodal sarcasm detection", Neurocomputing, 2025 Publication	<1%

9	"Proceeding of the 2nd International Conference on Machine Intelligence and Emerging Technologies", Springer Science and Business Media LLC, 2025 Publication	<1 %
10	"Advances and Trends in Artificial Intelligence. Theory and Applications", Springer Science and Business Media LLC, 2026 Publication	<1 %
11	Submitted to University of Sydney Student Paper	<1 %
12	"Combating Online Hostile Posts in Regional Languages during Emergency Situation", Springer Science and Business Media LLC, 2021 Publication	<1 %
13	Behnam Nikbakhtbideh, Linda Duffett-Leger, Mohammad Moshirpour. "Behavior Analysis of Parent-Child Interactions from Text", 2023 International Conference on Machine Learning and Applications (ICMLA), 2023 Publication	<1 %
14	Submitted to Vels University Student Paper	<1 %
15	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security", CRC Press, 2023 Publication	<1 %
16	repository.iutoic-dhaka.edu Internet Source	<1 %

17 "AI Technologies for Information Systems and Management Science", Springer Science and Business Media LLC, 2025 <1 %
Publication

18 "Advances in Artificial Intelligence and Machine Learning", Springer Science and Business Media LLC, 2025 <1 %
Publication

19 Anilkumar M. Bhadgale, M. Maheswari. "A Multi Model Deep Learning Framework for Sarcasm Detection in Regional Social Media: An Overview", 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET), 2025 <1 %
Publication

20 Arvind Dagur, Sohit Agarwal, Dharendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and Sustainable Innovation - Volume 3", CRC Press, 2026 <1 %
Publication

21 Tongguan Wang, Junkai Li, Guixin Su, Yongcheng Zhang, Dongyu Su, Yuxue Hu, Ying Sha. "RCLMuFN: Relational context learning and multiplex fusion network for multimodal sarcasm detection", Knowledge-Based Systems, 2025 <1 %
Publication

22 "Data Security and Privacy Protection", Springer Science and Business Media LLC, 2025 <1 %
Publication

23	C H. Raga Madhuri, Jaya Sankar Krishna Bandaru, Medisetti Srinu, Gangadhari Midhun Anand Vardhan. "AI-Powered Mental Health Screening and Support for Homeless Children", 2025 AI-Driven Smart Healthcare for Society 5.0, 2025 Publication	<1 %
24	www.mdpi.com Internet Source	<1 %
25	"Cognitive Computing – ICCC 2023", Springer Science and Business Media LLC, 2024 Publication	<1 %
26	Submitted to Dhirubhai Ambani University Student Paper	<1 %
27	Jaskaran Singh, Meenu Gupta, Rakesh Kumar. "Smart Technologies and Intelligent Computing", CRC Press, 2026 Publication	<1 %
28	"Computing and Machine Learning", Springer Science and Business Media LLC, 2026 Publication	<1 %
29	Jeremy Chang, Chia-Yun Hsu, Chung-Hsien Wu. "Detecting Sarcasm in Multiple Modalities by Leveraging Sarcasm Type Correlations", IEEE Access, 2026 Publication	<1 %
30	Lisong Ou, Zhixin Li. "Multi-modal Sarcasm Detection on Social Media via Multi-Granularity Information Fusion", ACM Transactions on Multimedia Computing, Communications, and Applications, 2025	<1 %

- 31 Mochamad Alfian Rosid, Daniel Oranova Siahaan, Ahmad Saikhu. "Sarcasm Detection in Indonesian-English Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi-Directional GRU", IEEE Access, 2024

Publication

- 32 Submitted to National Textile University, Faisalabad

Student Paper

- 33 Sunchan Park, Hyung Soon Kim. "Dimension-Specific Margins and Element-Wise Gradient Scaling for Enhanced Matryoshka Speaker Embedding", IEEE Access, 2025

Publication

- 34 Wenxuan Sun, Yanbin Guo, Shuwei Lin, Guoping Wang, Youpeng Zhu, Yao Zhang, Ye Du, Zibo Feng. "Assessment of the severity of lower extremity peripheral arterial disease based on gastrocnemius surface electromyography and mechanomyography signals", Medical & Biological Engineering & Computing, 2026

Publication

- 35 Xiaozhi Du, Shiming Zhang, Yanrong Zhou, Hongyuan Du. "A vulnerability severity prediction method based on bimodal data and multi-task learning", Journal of Systems and Software, 2024

Publication

- 36 artemis.cslab.ece.ntua.gr:8080

Internet Source

37	mkscienceset.com Internet Source	<1 %
38	scholar.archive.org Internet Source	<1 %
39	www.mecs-press.org Internet Source	<1 %
40	Submitted to British University in Egypt Student Paper	<1 %
41	Submitted to Delhi Technological University Student Paper	<1 %
42	Sathwik Acharya, Bhaskarjyoti Das, TSB Sudarshan. "Capturing the Concept Projection in Metaphorical Memes for Downstream Learning Tasks", IEEE Access, 2023 Publication	<1 %
43	lirias.kuleuven.be Internet Source	<1 %

Exclude quotes On

Exclude matches < 12 words

Exclude bibliography On