International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 662-678
https://doi.org/10.70917/ijcisim-2026-5074
Article

Enhancing Transformers with Polarity Encoders to Detect Sarcasm

S. Nagini[1], Karnam Akhil[2], Harshitha Upadhyayula[3], Tejasree Lokireddy[4], Bharath Siripuram[5], Thoran Chandhra Muvvala[6]*

[1]–[6]Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Bachupally, Hyderabad, Telangana, India.
[1] Email: nagini_s@vnrvjiet.in
[2] Email: akhil_k@vnrvjiet.in
[3] Email: harshitha.srikk@gmail.com
[4] Email: tejasreelokireddy@gmail.com
[5] Email: bharathbharath6456@gmail.com
[6] Email: thoranmuvvala@gmail.com

Abstract: Sarcasm detection is a complex task in natural language processing because it depends on implicit mood variations, contextual comparison, and the congruence of polarity between the literal form of expression and its intended meaning. Although the transformer-based models, including BERT and DeBERTa, have been taking a great leap in performance by introducing contextual self-attention, they mainly learn sarcasm patterns with an implicit hypothesis of sentiment polarity contradictions, which define the discourse of sarcasm. In this paper, a polarity-sensitive transformer model that explicitly incorporates sentiment data in representation learning is presented to detect sarcasm. In contrast to traditional fine-tuning methods, which treat sarcasm as a generic classification problem, the methodology adds sentiment-polarity cues to the embedding space, enabling the model to fine-tune contextual representations in a polarity-sensitive way. The proposed method will improve the ability of semantic representations of a context to reflect incongruity patterns in contextual segments. The positive results of experiments on benchmark sarcasm datasets indicate that explicit polarity integration is more robust and generalizes better than traditional transformer baselines, particularly in context-specific situations. The findings indicate that embedding-level sentiment improvement offers a sound theoretical and practical guideline in the process of expanding sarcasm detection beyond implicit contextual modeling.
Keywords: Sarcasm Detection, Transformer, DeBERTa, Sentiment Analysis, Context Modeling, Polarity Modeling, Text Classification, Natural Language Processing

1. INTRODUCTION

The development of digital communication has radically changed the way in which people share their views, feelings, and perceptions. Online review systems, social media sites, discussion boards, and news commentary paper are producing massive amounts of user-generated content daily. Figurative language, especially sarcasm, has become a common figure of speech in this ecosystem [1].

A. Background and Motivation

Sarcasm is normally used to represent meaning that is opposite to the actual implication of the words that are being said with a tendency of mocking, saying something negative or just trying to underpin a statement. Humans can typically identify sarcastic meaning by looking at the tone of sentences, but these types of expressions are not easily comprehended by computerized systems. As a result, the sarcasm detection problem has become an interesting issue of Natural Language Processing (NLP) [1].

Sarcasm detection is complex in nature. Sarcastic expressions depend majorly on the context of the situation, which is one of the main challenges [2]. This can be based on the context, common knowledge, previous conversation or cultural allusion. In many situations, interpretation of sarcasm involves identifying the incompatibility between what is expected and what is being said.

Domain variability is another major challenge. The use of sarcasm varies among social media comments like product reviews, and news headlines. Slang, abbreviations, emojis, and unconventional grammar are common features of informal conversational text and bring noise and ambiguity to the data. Conversely, intensive and contextually contingent irony in form of structured text like news headlines may be used [3].

Initial work in sarcasm detection was dependent on conventional machine learning methods. These methods used to be characterized by manual feature engineering, which added lexical, syntactic and sentiment-based features to a SVM, Naive Bayes model or Logistic Regression [4]. The introduction of deep learning led to an increase in the research of sarcasm detection by a huge margin. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) models enabled automatic extraction of features and improved modelling of context [5]. These methods reduced the use of manual design of features, and they performed superiorly compared to classical methods. However, sequential models were still criticized due to the demand to model long-range interactions and more complex context interactions especially in long conversations.

However, there are several research problems. Generalization to another domain is an urgent problem, and it is possible that a model trained on a dataset may not be able to generalize to another dataset since the language and style are not similar. These are just a few characteristics of a dataset, and it is possible that such characteristics can have a huge impact. Although the use of transformer architecture is beneficial in representation learning, it is possible that the quality and diversity of data can impact the performance [6]. All these indicate the need for a systematic implementation and assessment of various datasets.

B. Objectives

The main aim is to measure the effectiveness of detecting sarcastic phrases in different textual situations with modern transformer-based language modelling methods. The proposed methodology aims to present information regarding the performance and ability of contextual language models to detect sarcasm by examining performance trends and domain specific behaviour [7].

The main objectives of this research work are as follows:

- To implement and evaluate transformer-based sarcasm detection models, including BERT, RoBERTa, and DeBERTa, on sarcasm datasets.
- To enhance sarcasm classification performance using linguistic preprocessing strategies like augmenting synonyms in WordNet and addressing the problem of class imbalance with weighted loss.
- To hypothesize a polarity-aware encoding approach to incorporate explicitly sentiment polarity information into the transformer embedding space to better predict sarcasm incongruity.
- To perform cross domain evaluation on the Amazon product reviews and News Headlines data sets to determine the generalization ability.

A polarity-aware sarcasm detection framework is proposed, which directly incorporates sentiment information into transformer-based embeddings. Unlike other models, which treat sarcasm detection as a classification problem, sentiment polarity is directly incorporated at the representation level, which helps in better detecting incongruity-based patterns. In addition, data augmentation and class imbalance handling are also applied to improve model performance on different domains. The proposed model is evaluated on various benchmark datasets, i.e., Amazon Product Reviews and News Headlines datasets [8].

C. Acronyms

TABLE 1 Acronyms

Abbreviation	***Full Form***
NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
SVM	Support Vector Machine
NB	Naïve Bayes
LR	Logistic Regression
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
GRU	Gated Recurrent Unit
BiGRU	Bidirectional Gated Recurrent Unit
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach
DeBERTa	Decoding-enhanced BERT with Disentangled Attention
CLS	Classification Token
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
GPU	Graphics Processing Unit
CPU	Central Processing Unit
CUDA	Computer Unified Device Architecture

D. Mathematical Symbols and their Descriptions

TABLE 2 Mathematical Symbols and its Descriptions

Symbol	***Description***
$D = \{(x_i, y_i)\}_{(i=1)}^{N}$	Dataset consisting of text samples xi and their corresponding labels yi
xi	Input text instance (for example, an Amazon review or a news headline)
yi ∈ {0,1}	Ground truth label (0 indicates non-sarcastic, 1 indicates sarcastic)
Symbol	***Description***
N	Total number of samples in the dataset
wj	Word or token at position j in the input sentence.

W'j	Synonym-replaced token generated using WordNet augmentation
Synset(wj)	Set of WordNet synonyms corresponding to token wj
wc	Class weight assigned to class ccc for handling imbalance
Nc	Number of training samples belonging to class c
L	Loss function used during training
ŷi = σ(zi)	Predicted probability output for sample i
$\sigma(z) = 1/(1 + e^{(-z)})$	Sigmoid activation function
Ei ∈ Rd	Embedding vector representation of token i
Econtent	Content embedding component used in DeBERTa
Eposition	Position embedding component used in DeBERTa
E'i = Ei + λpi	Polarity-enhanced embedding representation of token i
si ∈ [-1,1]	Sentiment polarity score of tokens i, ranging from −1 to 1
Pi = WpSi	Projected polarity embedding vector for token i
Wp ∈ Rd×l	Trainable projection matrix for polarity embedding
λ	Learnable scaling factor that controls polarity contribution
Q,K,V ∈ Rn×d	Query, Key, and Value matrices used in self-attention
Qc, Kc	Content-based query and key matrices in DeBERTa attention
Qp, Kp	Position-based query and key matrices in DeBERTa attention
d	Embedding dimension of the transformer model
hCLS	Final contextual representation of the [CLS] token
ŷ = σ(W hCLS + b)	Output of the classification head
W ∈ Rl×d, b ∈ R	Trainable weights and bias of the classifier
Øi	SHAP attribution score for token or feature i
F	Set of all input features or tokens
S ⊆ F	Subset of features used in SHAP computation
f(S)	Model prediction using subset SSS of features

E. Paper Structure

Section II illustrates the related works on classification in text-based datasets and sarcasm corpuses, Section III shares the methodology of the overall paper, Section IV shows the results of implementing the proposed methodology and Section V and VI show the conclusion and any future works required. Table 1 illustrates acronyms and Table II highlights the symbol notations.

2. RELATED WORK

A. Traditional Sarcasm Detection Approaches

The original steps of sarcasm detection were based on manually created linguistic features and traditional machine learning algorithms. The use of sarcasm can be described as the contrast between the literal meaning and the implicated meaning, either by way of polarity inversion or incongruity in context [9]. Early methods tried to reflect these properties using features that were designed manually like mismatches in sentiment, punctuations, and contextual indicators. Thorough surveys like Joshi et al. (2017) suggest that in the initial period of sarcasm detection, much emphasis was placed on rule-based systems and statistical classifiers that worked on the lexical space [1].

Strategies of feature engineering were researched to accurately detect sarcastic phrases. A k-means based cluster formation method was introduced. The results showed the importance of features in improving the performance of the system for sarcasm detection based on product reviews and classification [9]. Similar studies were conducted to identify sarcastic sentences by applying sentiment dictionaries, emoticons, and syntactic features.

Conventional ML models were tested extensively for sarcasm detection. In the context of a Twitter-based sarcasm dataset [10], a comparative analysis of various conventional classifiers on the dataset was conducted. In their study, they applied a set of conventional classifiers such as NB, Decision Trees, Random Forest, k-Nearest Neighbours, LT, and SVMs. According to the results of the analysis, the performance of SVM was the best among all classifiers, with an F1-score of nearly 0.84 [1]. Similar trends were observed in other studies related to the detection of sarcasm in news headlines, where the performance of SVM models based on well-designed features was found to be nearly 90%.

Despite these achievements, these methods have their drawbacks. Domain and manual tuning are needed for handcrafted features, and these often do not work in different language area or styles such as the subtle irony in news headlines compared to the informality in social media. This shows the need to develop better representation techniques.

B. Deep Learning-Based Approaches

DL brought automatic methods of feature learning, greatly enhancing the performance of various NLP tasks. CNNs and RNNs were the neural architectures that allowed models to learn semantic and syntactic patterns directly off the input without prior feature engineering.

It was shown by Ghosh and Veale (2016) that convolutional architectures had the capacity to capture local contextual patterns in sarcastic tweets because they are learned to represent phrases. They demonstrated their model to be better at performance compared to traditional machine learning strategies because they were able to identify the lexical structures of sarcasm [5].

Recurrent neural networks, especially LSTM networks, performed better at modelling sequential dependencies in text to improve sarcasm detection. Amir et al. (2016) suggested a structure based on LSTM and integrated with conversational setting to identify sarcasm in online conversations. The model managed to model contextual incongruities and normally suggest sarcastic intentions by considering the context of the response and the preceding conversation. Similarly, BiLSTM models allowed a model to have access to the bi-directional context information, and this allowed identification of fine semantic differences [2].

The hybrid architectures including CNN and LSTM were also considered to understand the local and global context. Kulkarni and Rekha proposed a deep neural network, which combined word representations with convolutional networks and BiLSTM to forecast sarcasm as polarity inversion task [11]. They have trained their system on cases of literal and contextual discrepancy between sentence meaning and context meaning and obtained accuracy rates of around 88 percent in datasets of social media and news headlines.

The use of pretrained word embedding models like Word2Vec and GloVe was an important development [12]. These embeddings gave dense representations in vectors of semantic relations between words. In their article, Khatri and Pranav (2020) studied how sarcasm can be detected using GloVe and BERT embeddings on a Twitter-based

dataset. They reported that a logistic regression classifier on GloVe sentence embeddings performed slightly better than the corresponding classifier on BERT embeddings with F1-scores of about 0.69 and 0.63 respectively [13]. The authors indicated that contextual embeddings can be affected by the size of the given dataset and the properties of the domain. Moreover, they also found that the integration of conversational context and the target response enhanced better classification, which supports the role of contextual information with sarcasm detection [14].

DL models greatly enhanced the representation learning, but the models remained inadequate in long-range dependency and intricate semantic interaction. These complexities encouraged the use of transformer-based architectures that were able to represent contextual relationships in text at a better level.

C. Sarcasm Detection based on Transformers.

Transformer-based architectures have emerged as the new paradigm of sarcasm detection because they can simulate the contextual dependencies with self-attention. BERT, RoBERTa, and DeBERTa generate contextualized representations of tokens which can be modified for classification problems.

Transformer-based models are effective in the sarcasm detection in various studies. Baruah et al. (2020) experimented with various architectures using Twitter and Reddit sarcasm data and found that a fine-tuned BERT classifier achieved a much higher score of approximately 0.74 on Twitter data with an F1-score of 0.74 compared to BiLSTM and SVM baselines. These findings validated the presence of delicate semantic connections in contextualized embeddings which turned out to be involved to depict with the previous models [3].

In the Sarcasm Detection Shared Task during the Workshop on Figurative Language Processing, it was shown that transformer models were superior compared to conventional models. The best models were the optional attention mechanism-based models like BERT or RoBERTa, or an ensemble. A combination of BERT embeddings, BiLSTM layers, and multi-view attention is the model that has demonstrated the best performance with a F1-score of 0.93, thus demonstrating the effectiveness of transformer-based contextual representations.

Different transformer-based architectures were tested for text classification. BERT and DistilBERT, RoBERTa and BiGRU were tested by Jayaraman et al. on a news headline sarcasm dataset. Their findings showed that the RoBERTa model was the most accurate in classification both compared to traditional machine learning paradigms and previous neural network frameworks. Equally, Potamias et al. have shown that transformer-based models that are always superior to CNN and RNN baselines in various sarcasm detection datasets [15].

Transformer models, despite the high performance, need large training datasets and heavy computational power [16]. Moreover, self-attention mechanisms encode contextual relationships but fail to encode sarcasm-specific linguistic information, i.e. polarity contradictions or semantic incongruity [17]. Investigation has shifted to hybrid architectures, which use extra linguistic signals as well as encodings in transformers.

D. Sentiment-Aware and Context-aware Sarcasm Detection.

Recent studies have given attention to incorporation of the sentiment knowledge and contextual evidence into sarcasm detection models. Shangipour Atai et al. recommended an architecture that is sentiment-aware by integrating BERT embeddings with aspect-based sentiment analysis [18]. The sentiment signals of elements in conversation and response text are extracted in their model, allowing the network to learn contrastive sentiment relationships which imply sarcasm.

The authors noted that sarcasm usually occurs when the feeling conveyed in a reply is a contradiction of the feeling of the context that precedes it [19]. Their explicit modelling of this relationship gained greater success on benchmark datasets of Twitter and Reddit sarcasm. This piece of work shows the relevance of identifying sentiment incongruity, which is one of the linguistic features of a sarcastic speech.

There have been other investigations on contextual modelling approaches to sarcasm detection. For instance, multi-turn dialogue models use conversational history to get a clearer idea of the pragmatic meaning of a response.

These models indicate that the accuracy of sarcasm detection can be better, when the information about a context is added and especially in the context of social media where sarcasm is usually contingent on conversational indicators.

Techniques such as ensemble learning and data augmentation promote model generalization. In order to augment the training data, research has shifted to applied lexical augmentation methods. Ensemble learning combines multiple transformer models, and these have performed better in sarcasm detection compared to conventional methods and transformer models.

E. Research Gap

There are still several shortcomings in the current research. The gaps that have been identified in the previous studies are as follows:

- Existing transformer-based methods are highly dependent on implicit learning and fail to explicitly represent significant linguistic characteristics of sarcasm, including polarity reversal and semantic incongruity, so they are less effective at understanding more subtle expressions of sarcasm.
- Sentiment information is not strongly captured as part of the representation learning algorithm, and current approaches do not tend to be sensitive to token-level sentiment differences.
- Conventional models demonstrate poor cross-domain generalization and do not have strong and linguistically driven data augmentation methods, and offer little interpretability, making them less useful in real world scenarios.

These gaps build the motivation to the proposed polarity-sensitive transformer framework, incorporating sentiment polarity data into representation learning that is robust in augmentation and imbalance management and more interpretable with explainable AI solutions.

3. METHODOLOGY

A. Problem statement

Sarcasm detection is a complex NLP task as sarcastic expressions often convey an intended meaning that contradicts their representation. This contradiction is commonly reflected through polarity reversal, where positive lexical cues are used to express negative sentiment or criticism, making sarcasm complex to identify using standard sentiment analysis or conventional text classification approaches. Although transformer-based architectures such as BERT, RoBERTa, and DeBERTa achieve strong contextual representation learning, they primarily rely on implicit learning of sarcasm cues and do not explicitly model polarity incongruity within the embedding and encoding pipeline. Sarcasm data commonly has domain variation and category. imbalance, which may have an adverse impact on generalization between various text sources. such as product reviews and news headlines. Thus, the issue discussed in the task to be completed in this work is the design and implementation of a transformer-based sarcasm detector. framework which explicitly incorporates sentiment polarity information into. representation learning enhances robustness with controlled augmentation. and imbalance management and delivers a steady performance on a variety of. benchmark datasets.

Equation 1 shows the dataset where x and y belong to binary classes:

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad y_i \in \{0,1\} \tag{1}$$

Training a model to predict as shown in Equation 2:

$$\hat{y}_i = \sigma(W h_{CLS} + b) \tag{2}$$

where each tokens representation includes sentiment polarity as shown in Equation 3:

$$E'_i = E_i + \lambda p_i, p_i = W_{psi} \tag{3}$$

so that the model captures polarity incongruity and classification performance increases across datasets.

B. Datasets

Two popular benchmark datasets are used as the research material of this paper. on the detection of sarcasm as experimental evidence. The initial data is Amazon Product Reviews data, and this data consists of user product reviews. that have been annotated using sarcastic and non-sarcastic words. This data set will be used to document informal language usage, including tone, use of language and contextual forms. The dataset contains 8,000 samples, 4,000 and There are 4,000 sarcastic and non-sarcastic, respectively [9].

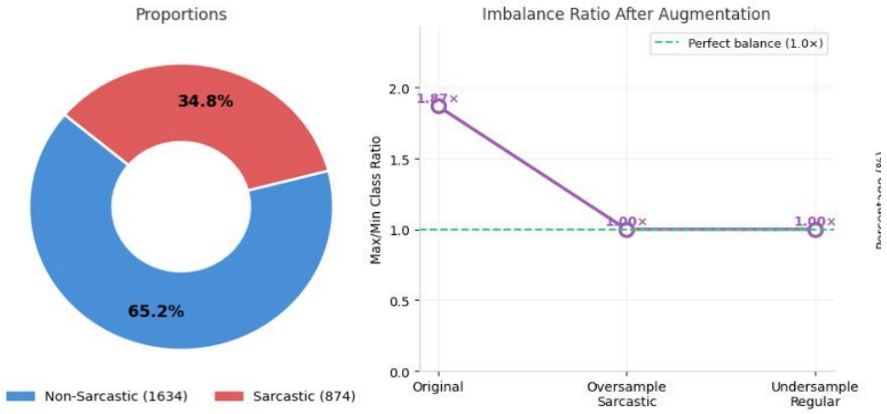


Figure 1: Class Imbalance in Amazon Review Dataset and ratio after Augmentation

Figure 1 shows the class distribution of the amazon review dataset and the effect of data augmentation on the dataset to achieve balanced ration of classes explaining the role of over and under sampling to create a non-skewed dataset.

Figure 2: Sentence length distribution of Amazon Review Dataset

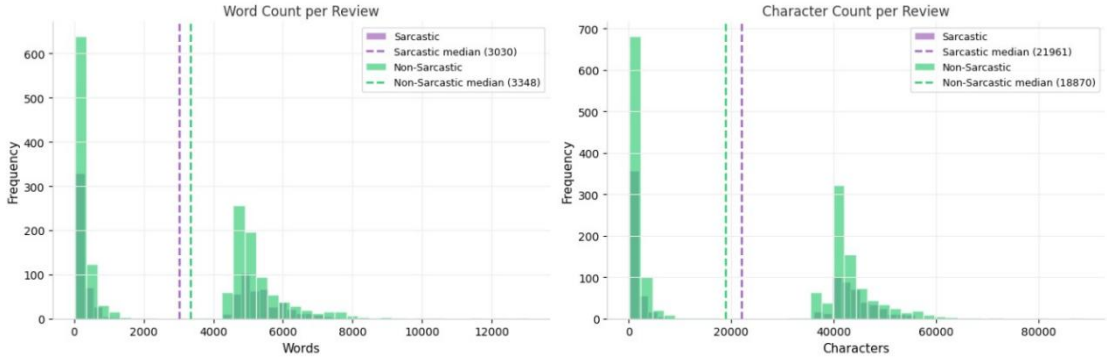


Figure 2 shows the length of texts in sarcastic and non-sarcastic samples using word count and character count histograms. The non-sarcastic cases have a little longer median length than sarcastic cases, suggesting that the non-sarcastic ones are longer and more descriptive. Conversely, sarcastic texts are comparatively shorter with many being implicitly expressed through shorter constructions. These observations support the application of models that utilize transformer-based features that can capture contextual features other than mere length-based features.

The second dataset that will be used in this paper is the News Headings dataset on identifying sarcasm. This information has structured news headlines classified as either sarcastic or non-sarcastic which provides it with an opposing domain in which the linguistic patterns are more formal [17]. The data set contains 26, 709 samples, each of which contains 11, 748 sarcastic and 14,961 non-sarcastic.

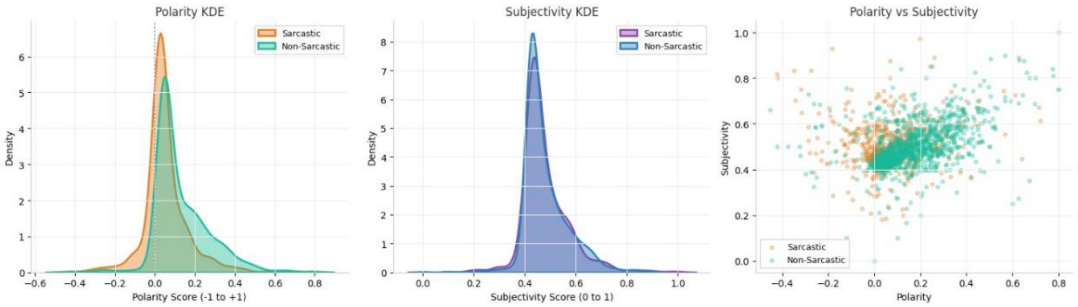


Figure 3: Sentiment Polarity and Subjectivity Distribution on Amazon Reviews Dataset

Figure 3 presents the distribution of sentiment polarity and subjectivity for sarcastic and non-sarcastic texts, along with their relationship. From the polarity KDE plot, it is observed that sarcastic samples are denser around neutral to slightly negative values, whereas non-sarcastic texts show a wider spread toward positive polarity. This indicates that sarcasm often involves subtle or implicit sentiment rather than strongly expressed polarity. The subjectivity distribution shows significant overlap between both classes, suggesting that both sarcastic and non-sarcastic texts are generally subjective in nature, making subjectivity alone insufficient for discrimination. These observations reinforce the need for polarity-aware and context-sensitive models, as simple sentiment signals are not adequate to accurately detect sarcasm.

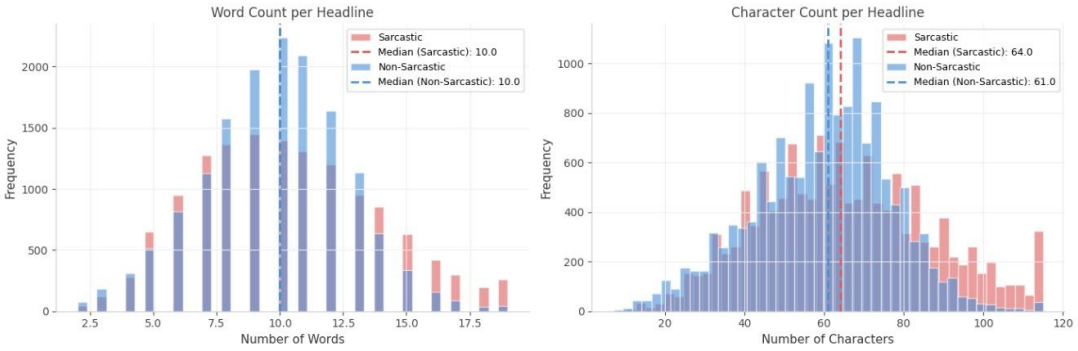


Figure 4: Headline length distribution of News Headlines Dataset

Figure 4 illustrates the distribution of headline lengths of sarcastic and non-sarcastic samples in the news headlines dataset is shown in Fig 4 both in the number of words and in the number of characters. The word count distribution indicates a significant overlap between the two classes with a median of around 10 words in each, so that the length in terms of words of headline is not significantly differentiating between sarcastic and non-sarcastic content. Nevertheless, the distribution of the number of characters indicates that there is a slight difference, with sarcastic headlines having slightly higher median length than non-sarcastic ones. This implies that, although the quantity of words used will be comparable in number, sarcastic headlines can be longer and more expressive with words to depict implicit meaning. The high coincidence of both distributions suggests that the length of the headline is not a good characteristic to use when detecting sarcasm, which is why contextual and semantic modelling are crucial in making the correct classification.

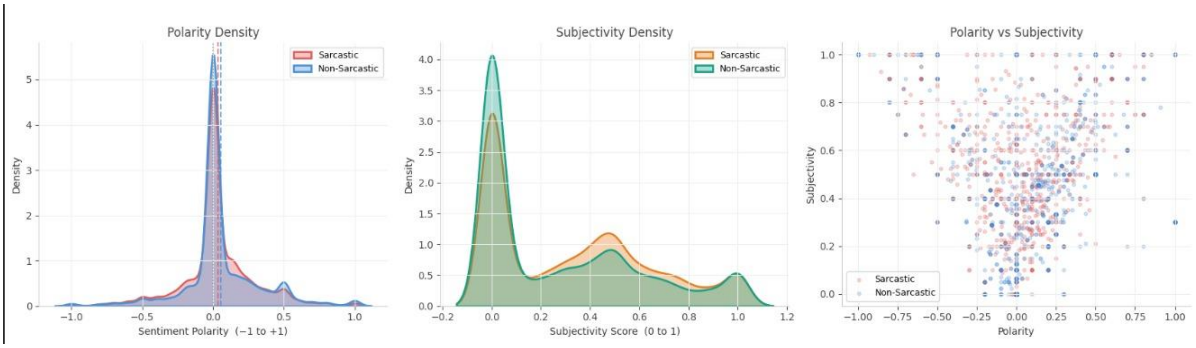
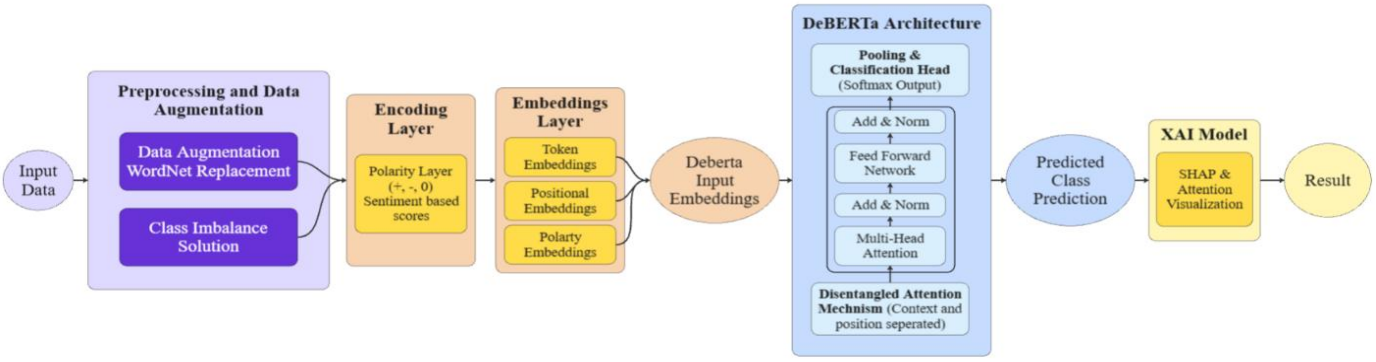


Figure 5: Sentiment Polarity and Subjectivity Distribution on News Headlines Dataset

Figure 5 shows distribution of sentiment polarity and subjectivity for sarcastic and non-sarcastic samples in the news headlines dataset. Both classes are concentrated around neutral polarity values with significant overlap, indicating that headlines generally express minimal explicit sentiment. The subjectivity distribution suggests moderate subjectivity with slight variation in sarcastic instances, but no clear distinction between classes. The scatter plot further confirms this overlap, with both categories clustered around neutral polarity and mid-range subjectivity, making sarcasm difficult to distinguish using sentiment alone. In comparison to the Amazon dataset, which exhibited a wider polarity spread and clearer separation between classes, the news headlines dataset demonstrates more subtle and implicit expressions of sarcasm.

Table 3 Dataset Description

Dataset	***Domain***	***Samples***	***Sarcastic***	***Non-Sarcastic***
Amazon Product Reviews	User-generated reviews	8,000	4,000	4,000
News Headlines Dataset	News headlines	26,709	11,748	14,961



As shown in Table 3, the Amazon Reviews dataset is balanced, containing an equal number of sarcastic and non-sarcastic samples. In contrast, the News Headlines dataset exhibits a mild class imbalance, making it suitable for evaluating model robustness under real-world distribution shifts. The use of both datasets enables cross-domain evaluation, covering informal opinion-based sarcasm in reviews as well as more structured irony patterns present in news headlines.

C. Overview of the Proposed Framework

As shown in Figure 6, the proposed sarcasm detection framework is a multi-stage classification architecture that consists of

1. Preprocessing and data augmentation for class imbalance,
2. Polarity-aware encoding,
3. Enhanced transformer-based feature extraction with DeBERTa, and
4. An explainability module to interpret models.

A standard DeBERTa fine-tuning for binary classifications is used. It is suggested that the application of polarity-sensitive encoding before computing.

Data Preprocessing and Augmentation

1. Text Normalization

Input text samples are normalized through lowercasing, removal of extraneous whitespace, and token standardization. For a given input sentence, a set of non-stopword tokens are randomly replaced with semantically similar synonym tokens. Let equation 1 represent the dataset

Figure 6: Architecture of the polarity-based BERT model for sarcasm detection

2. Data Augmentation using WordNet Replacement

To enhance generalization and mitigate overfitting, synonym-based augmentation is applied using WordNet lexical replacement. For each input sentence, a subset of non-stop word tokens is randomly selected and replaced with semantically similar synonyms. Given a sentence x = {w1, w2, …, wn}, augmented samples x′ are generated by Equation 4:

$$w'j \sim \text{Synset}(wj) \quad (4)$$

where wj′ is sampled from WordNet's synonym set of wj, maintaining grammatical structure. The use of lexical variation that still conveys meaning can be beneficial in sarcasm detection because it can have significant variation in surface form.

3. Handling Class Imbalance

Class imbalance is common to sarcasm datasets. In order to solve this, weighted loss functions are used. Class weights can be defined as shown in Equation 5:

$$wc = N/2Nc \quad (5)$$

In which Nc is the number of samples in class c. The weighted binary cross-entropy loss is obtained as:

$$L = -\sum \quad (i = 1)^{N} w_{(y_i)}[y_i lo(\hat{y}_i) + (1 - y_i) lo(1 - \hat{y}_i)) \quad (6)$$

Equation 6 illustrates equal contributions of the two classes in terms of gradient.

Baseline Model: Standard DeBERTa Fine-Tuning

The basic architecture takes the DeBERTa as a pretrained transformer encoder and then a linear classification head. The DeBERTa (Decoding-enhanced BERT with Disentangled Attention) has two major improvements on BERT:

1. Disentangled attention mechanism
2. Enhanced mask decoding strategy

DeBERTa is different, unlike BERT, which uses a single embedding vector in place of representations of tokens, it uses content embedding and position embedding

In the case of disentangled attention, the attention weights are calculated through individual interactions. It is shown in Equation 7:

Attention = SoftMax $Q_c K_c^T + Q_c K_p^T + Q_p K_c^T$ (7)

where:

- Qc, Kc represent content-based queries and keys,
- Qp, Kp represent position-based queries and keys.

This division enhances the contextual modelling, especially the syntactic and semantic alignment.

Polarity-Based Encoder

Sarcasm tends to engage positive surface lexical and negative contextual implication. A polarity-aware encoding mechanism is proposed to capture this phenomenon. The sentiment polarity si is for each token ti calculated on a lexicon-based or pretrained sentiment analyser. The score of polarity is normalized to: si ∈ [-1,1]

A learner projection is formed with the help of a projection to an auxiliary polarity represented as Equation 8:

pi = WpSi (8)

where Wp ∈ Rd×k projects scalar polarity into embedding dimension.

D. Algorithm

Algorithm1: DeBERTa with Polarity-Aware Encoding for Sarcasm Detection.

Input: Text dataset D = (x_i, y_i) where x_i is input text and $y_i \in \{0, 1\}$

Output: A trained polarity-aware transformer model that predicts the sarcasm label ŷi ∈ {0, 1} for a given input:

4. : Preprocessing & Augmentation

For each input xi = {wj}, generate augmented tokens:

$wj' \sim Synset(wj)$

5. : Polarity-Aware Embedding

Each token embedding is enhanced using sentiment polarity:

Ej′ = Ej + λWpSj

6. : Contextual Encoding

Enhanced embeddings are passed through DeBERTa:

hCLS = DeBERTa(E′)

7. : Prediction

ŷi = σ(WhCLS + b)

8. : Training Objective

Model parameters are optimized using weighted binary cross-entropy:

min L (yi, ŷi)

The proposed algorithm 1 integrates polarity-aware embedding with DeBERTa to explicitly model sentiment incongruity in text, a key indicator of sarcasm. By combining data augmentation, sentiment-enhanced representations, and transformer-based contextual encoding, the model improves sarcasm detection performance.

E. Experimental Setup

The proposed models are implemented using the Hugging Face Transformers library with PyTorch backend. Training is performed using the AdamW optimizer with a learning rate of 2×10^−5. The batch size is fixed at 8, and models are trained for up to 10 epochs depending on convergence behaviour. The maximum length of input sequence is 512 tokens to provide adequate coverage of long review−based samples, and shorter inputs are padded and long ones truncated.

The datasets are divided into training and testing splits using an 80:20 ratio. Additionally, a validation subset is created from the training set to support early stopping and hyperparameter tuning. Early stopping is applied based on validation F1-score with a patience value of 2 epochs to prevent overfitting. Dropout regularization is retained as defined in the pretrained transformer backbone. All experiments are executed on a CUDA-enabled GPU when available; otherwise, training and evaluation are performed on CPU.

4. RESULTS

The effectiveness of the approach is determined in relation to existing techniques of sarcasm detection. Conventional machine learning models like SVMs have reached an approximate F1-score of 0.84 when working with Twitter-based data [4]. F1-scores of 0.85 to 0.87 have been improved with models of DL, such as CNN and LSTM structures [2][5].

Models that optimize transformer architecture like BERT achieved F1-score of 0.90. Comparatively, the proposed polarity-conscientious DeBERTa model has an F1-score of 0.9886 on the Amazon Reviews dataset, and 0.9426 on the News Headlines Dataset.

TABLE 4 Comparison of existing models

Model	Approach Type	Dataset	F1 Score
SVM [20]	Traditional ML	Twitter	0.84
CNN [21]	DL	Twitter	0.85
LSTM [22]	DL	Conversation	0.87
BERT [14]	Transformer	Twitter	0.69
RoBERTa	Transformer	News Headlines	0.90
DeBERTa	Transformer	Amazon	0.98
Proposed Model	Polarity-Aware Transformer	News Headlines	0.9426
Proposed Model	Polarity-Aware Transformer	Amazon	0.9886

As shown in Table 4, the proposed model achieves higher performance compared to existing approaches. The variations in the performances, however, can be attributed to the characteristics of the datasets, preprocessing

approaches, and experiment setups. The results indicate the effectiveness of polarity-aware embeddings in capturing sarcasm-specific patterns.

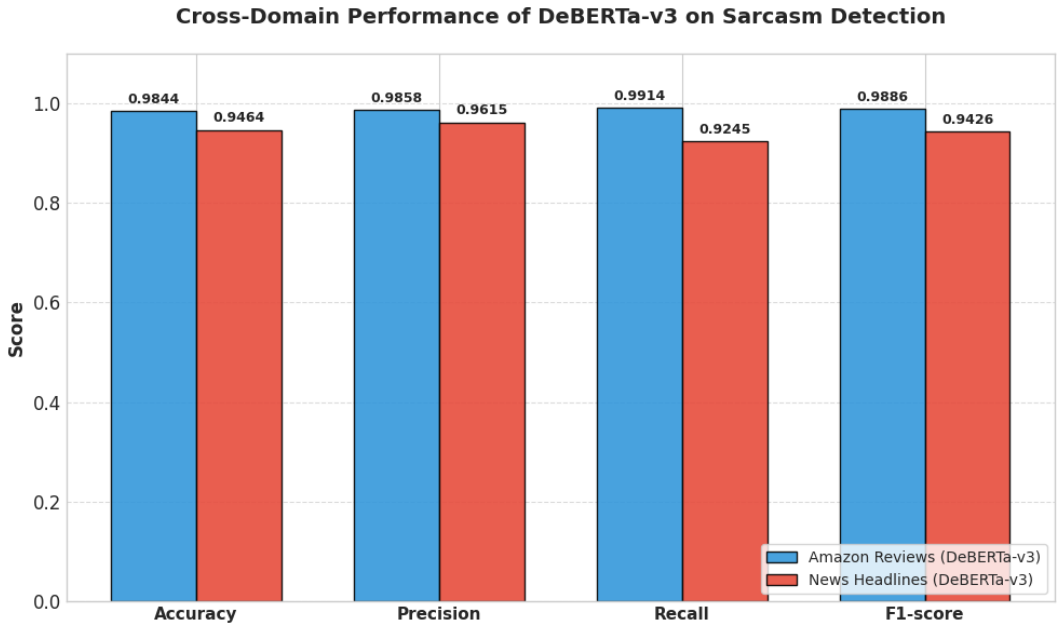


Figure 7: Cross-Domain Generalization

Figure 7 shows that the proposed DeBERTa-v3 sarcasm detection framework can be used to generalize across domains. Even though the model has nearly ideal performance on the Amazon Reviews dataset, there is a major decline on the News Headlines dataset on all evaluation criterion. However, the model is highly precise and has a high F1-score that proves the effectiveness of the proposed augmentation and polarity-aware representation learning method to enhance robustness.

Table 5 Performance of BERT Based Models on Amazon Review Dataset

Model	***Training Configuration***	***Accuracy***	***Precision***	***Recall***	***F1-score***
BERT (Base)	2 epochs	0.8327	0.9412	0.5517	0.6957
BERT (Base)	5 epochs	0.8406	0.8406	0.6667	0.7436
BERT (Base)	4 epochs + Weighted Loss	0.8526	0.8571	0.6897	0.7643
RoBERTa-base	5 epochs	0.8924	0.8947	0.7816	0.8344
RoBERTa-base	10 epochs	0.9044	0.8987	0.8161	0.8554
DeBERTa-v3	5 epochs + Simple Augmentation	0.9765	0.9701	0.9824	0.9811
DeBERTa-v3	5 epochs + WordNet Augmentation	0.9789	0.9738	0.9824	0.9830
DeBERTa-v3	10 epochs + WordNet + Weighted Loss	0.9800	0.9838	0.9894	0.9886

	+ Early Stopping				
DeBERTa-v3	10 epochs + WordNet + Weighted Loss + Polarity Embeddings	0.9844	0.9858	0.9914	0.9886

As shown in Table 5, the experimental results show a gradual and significant change in the improvement in the model architecture as the model architecture increases to BERT, to RoBERTa, and ultimately to DeBERTa-v3 with augmentation and imbalance handling strategies.

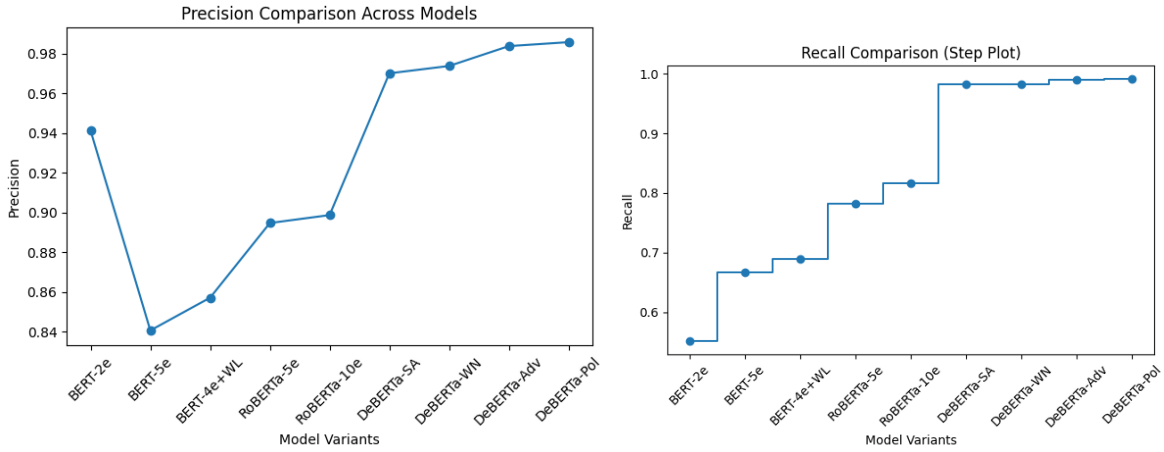


Figure 8: Model Precision and Recall Comparision

Figure 8 illustrates the variation in precision across different model configurations. The results highlight that incorporating augmentation strategies, weighted loss, and polarity-aware embeddings lead to consistent performance gains, with the highest precision obtained using the proposed polarity-enhanced DeBERTa model.

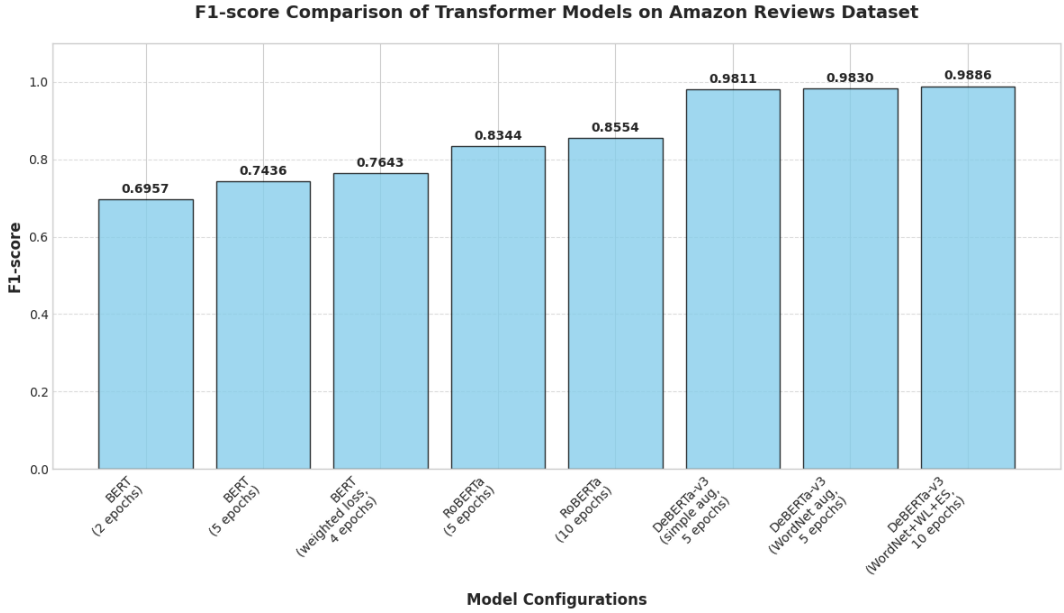


Figure 9: Model Performance Comparison (F1-score across Models)

As shown in Figure. 9, compared to BERT, RoBERTa-base demonstrates a significant improvement in recall and overall F1-score overall, which is understandable as the pretraining strategies and dynamic masking provided better results. Nevertheless, the highest gains can be noticed using DeBERTa-v3. With a basic implementation and without any further complexity, DeBERTa demonstrates an F1-score of over 0.98, which indicates that the power of disentangled attention processes to capture contextual dependencies. The use of WordNet based augmentation also increases performance meaning that semantically regulated lexical variation improves generalization. The last design, WordNet augmentation and weighted loss and early stopping, has the greatest performance (F1-score of 0.9886) on the Amazon dataset. In the News Headlines dataset, DeBERTa has a high level of generalization with an F1-score of 0.9426, which is consistent across domains.

5. CONCLUSION

This paper presented a framework of detecting sarcasm with polarity-conscious transformers, and the focus was on the explicit use of sentiment polarity in representation learning. In contrast to traditional methods of transformer fine-tuning based on implicit contextual learning, the proposed approach integrates polarity signals directly into the embedding space, which allows better modeling the sentiment incongruity, which is a fundamental aspect of sarcasm. The experimental analysis performed on the Amazon Product Reviews and news headlines datasets indicates that transformer models, specifically the DeBERTa model, have high performance, and data augmentation using WordNet and weighted loss are also beneficial in enhancing the robustness of the architecture, in terms of lexical diversity and reducing class imbalance. The findings show that the polarity-conscious encoding enhances the capability of the model to identify the subtle sentiment contradictions that are usually hard to detect with the use of the contextual embeddings only. Also, explainable algorithms like SHAP and attention visualization offer interpretability by highlighting token-level effects on prediction results, enhancing the transparency of the classification process. In general, the results indicate that the introduction of linguistically meaningful polarity cues into transformer networks is a promising approach to achieving a higher sarcasm detection rate and building more context-aware natural language understanding models.

6. FUTURE SCOPE

- Extending the proposed polarity-aware transformer framework to multi-turn conversational sarcasm detection, where sarcasm depends heavily on dialogue history.
- Investigating cross-lingual and multilingual sarcasm detection, particularly for low-resource languages where sarcastic patterns vary culturally.
- Incorporating external knowledge sources such as knowledge graphs or commonsense reasoning models to better handle sarcasm requiring real-world understanding.
- Improving explainability by integrating advanced XAI methods such as Integrated Gradients and LIME and performing user-based evaluation of explanation quality.

REFERENCES

1. . Joshi A, Bhattacharyya V, Carman MJ (2017) Automatic sarcasm detection: A survey. ACM Computing Surveys 50(5):1–22.
2. . Amir S, Wallace BC, Lyu H, Carvalho P, Silva MJ (2016) Modelling context with user embeddings for sarcasm detection in social media. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 167–177.
3. . Baruah A, Das K, Barbhuiya FA, Dey K (2020) Context-aware sarcasm detection using transformer-based models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 765–774.
4. . Siddiqui S, Chandra M, Kumar P (2021) Comparative evaluation of machine learning classifiers for sarcasm detection on Twitter datasets. International Journal of Advanced Computer Science and Applications 12(6):1–10.
5. . Ghosh A, Veale T (2016) Fracking sarcasm using neural network. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2016), 161–169.
6. . Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, 4171–4186.

7. . Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
8. . Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS 2017), 4765–4774.
9. . Kumar A, Harish BS (2018) Feature selection for sarcasm detection using machine learning techniques. International Journal of Engineering & Technology 7(4):567–572.
10. . Tay Y, Tuan LA, Hui SC (2018) Reasoning with sarcasm by reading in-between. In: Proceedings of the ACL Workshop on Figurative Language Processing, 67–76.
11. . Kulkarni V, Rekha B (2020) Deep neural network models for sarcasm detection in social media and news headlines. Procedia Computer Science 167:435–444. https://doi.org/10.1016/j.procs.2020.03.247
12. . Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
13. . Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1532–1543.
14. . Khatri J, Pranav P (2020) Sarcasm detection using BERT and GloVe embeddings. In: Proceedings of the International Conference on Computing and Communication Systems (ICCCS 2020), 208–213.
15. . Potamias R, Siolas G, Stafylopatis A (2020) A transformer-based approach for sarcasm detection. In: Proceedings of the European Conference on Artificial Intelligence (ECAI 2020), 2500–2507.
16. . He P, Liu X, Gao J, Chen W (2021) DeBERTa: Decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (ICLR 2021).
17. . Jayaraman S, et al. (2021) Performance comparison of transformer-based models for sarcasm detection. In: Proceedings of NLP Conference, 112–120.
18. . Wei J, Zou K (2019) EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 6382–6388.
19. . Shangipour Atai S, et al. (2021) Sentiment-aware sarcasm detection using BERT and aspect-based analysis. Journal of Artificial Intelligence Research 70:123–140.
20. . Chia, C., et al. (2018) Sarcasm detection using SVM-radial on Twitter datasets. *International Journal of Recent Technology and Engineering* 12(5)
21. . Darkunde, M. A., et al. (2020) Sarcasm detection using CNN and LSTM models on Twitter data. *International Journal of Scientific Development and Research*
22. . Cai, Y., et al. (2019) Sarcasm detection in multimodal conversational data using Bi-LSTM. *Proceedings of NLP Research*