

Incongruity-Aware Sarcasm Detection

Obydullah Sayed¹, M. Showaib Rahman Tanveer¹, M. Samirul Islam¹, Saiful Azad¹, Sudip Ghoshal¹, Shankha Shubhra Sarkar¹, and Tanpia Tasnim¹

Green University of Bangladesh, Purbachal American City, Dhaka, Bangladesh

Abstract. Sarcasm detection is difficult because intended meaning often contradicts literal wording, which misleads sentiment analysis and other NLP systems. Although transformer encoders such as RoBERTa capture context well, they can miss subtle contrastive cues. This paper proposes an *incongruity-aware* framework, IDL-RoBERTa, that injects an interpretable sentiment-contrast signal into a RoBERTa-based classifier. Using VADER, we compute sentence-level polarity and token-level polarity, derive an incongruity feature from their disagreement, and align it with the token sequence as an additional feature channel. IDL-RoBERTa is trained with a multi-task objective that combines sarcasm classification with auxiliary regression of sentence-level incongruity to encourage attention to contrast rather than surface lexical patterns. Model parameters are optimized using gradient-based learning (mini-batch gradient descent) with an AdamW optimizer and a warmup schedule, as configured in the final notebooks. Experiments on the News Headlines Sarcasm Dataset v2 and a Twitter sarcasm dataset show consistent improvements over classical CNN/LSTM baselines and a strong RoBERTa baseline in our recorded runs: headline test F1 rises from 0.8019 to 0.8966, and Twitter test F1 from 0.6541 to 0.7342. We also report back-translation ablations on Twitter to examine augmentation effects. Overall, the approach improves performance while preserving interpretability through an explicit, inspectable proxy for incongruity.

Keywords: Sarcasm Detection · Incongruity Theory · IDL-RoBERTa · Back Translation · Sentiment Analysis

1 Introduction

Sarcasm is a common linguistic phenomenon in online communication where the intended meaning contrasts with the literal surface form. This mismatch frequently leads to failures in sentiment analysis and downstream NLP applications such as opinion mining, customer feedback analysis, and content moderation. Modern transformer-based encoders such as RoBERTa [2] provide strong contextual representations, but sarcasm remains challenging because it often depends on implicit *incongruity* between sentiment and situation. This paper investigates **incongruity-aware sarcasm detection** by explicitly modeling sentiment-based incongruity features and integrating them into the proposed **IDL-RoBERTa** architecture. We implement and evaluate (i) a transformer

baseline (RoBERTa) and (ii) IDL-RoBERTa, which augments contextual token representations with incongruity signals computed using VADER [3]. As a secondary analysis, we also report an augmentation (BT) ablation on a Twitter sarcasm dataset [4]. Sarcasm is often described as a form of semantic or pragmatic contrast: positive words can express negative intent (or vice versa) depending on context [5]. As a result, purely lexical sentiment cues are insufficient, and models can overfit to superficial patterns.

Motivated by the growing need for robust sarcasm detection, this work investigates whether explicitly modeling sentiment incongruity can improve the identification of sarcastic news headlines. Specifically, we examine the contrast between positive and negative sentiment within contextual information, as such incongruity is widely recognized as a fundamental linguistic characteristic of sarcasm. To address these challenges, this paper presents a hybrid incongruity-aware sarcasm detection framework that integrates contextual semantic representations with sentiment-based linguistic cues. The main contributions of this work are summarized as follows:

- **A novel hybrid sarcasm detection framework** that integrates transformer-based contextual representations with token-level incongruity features derived from VADER sentiment through the proposed **IDL-RoBERTa** architecture. The framework further employs a multi-task learning formulation that jointly performs sarcasm classification and sentence-level incongruity regression to encourage the model to focus on contrastive linguistic cues.
- **A reproducible training and evaluation pipeline** implemented in Python notebooks for both news headline and Twitter datasets, incorporating mini-batch gradient-based optimization and carefully configured hyperparameters to maximize classification performance while ensuring reproducibility.
- **A comprehensive empirical evaluation** comprising ablation studies across Baseline, IDL-only, BT-only, and BT+IDL configurations on both news headline and Twitter datasets, supported by quantitative performance metrics and comparative visualizations to systematically analyze the contributions of incongruity features and data augmentation.

The remainder of the paper is organized as follows. Section 2 describes the relevant existing works. Section 3 elaborates the proposed sarcasm detection framework with workable example. Subsequently, Section 4 delves into the specifics of the performance comparison of the experimental results to validate the proposed model. Finally, in Section 5, concluding remarks and future research directions are stated.

2 Related Works

Early work in sarcasm detection focused on lexical cues, rule-based patterns, and hand-crafted features such as sentiment contrast, punctuation, and pragmatic markers. Recent studies have moved toward representation learning and transformer encoders that grasp contextual meaning. This section reviews (i) theoretical motivation for modeling incongruity, (ii) classical and neural approaches to

sarcasm detection, (iii) transformer-based sarcasm classification, (iv) sentiment resources such as VADER, and (v) data augmentation via back-translation.

Theoretical background: Sarcasm is generally characterized by the contrast between literal expression and intended meaning, often referred to as *sentiment incongruity*. A significant perspective characterizes sarcasm as the fusion of a positive sentiment expression with a negative situation, prompting clear modeling of sentiment disparity [5]. In application, incongruity may be estimated through sentiment analyzers, emotional lexicons, or contextual representations. Although sentiment-based methods may struggle with subtle or informal language, explicit incongruity features remain valuable as interpretable signals that complement deep contextual models.

Evolution of Sarcasm Detection Methods: Early sarcasm detection techniques depended on manually crafted features like n-grams, polarity changes, intensifiers, emoticons, hashtags, and syntactic indicators along with classifiers such as SVM and logistic regression [1]. While effective on constrained datasets, these approaches require extensive feature engineering and often generalize poorly across domains [1]. Deep learning techniques subsequently minimized manual feature engineering by acquiring semantic representations directly from text. CNNs and RNNs improved performance on short texts [13], but their constrained capacity to identify long-range contextual connections limits their efficacy in understanding intricate sarcasm [14].

Transformer Models for Sarcasm Detection: Transformer models such as BERT [6] and RoBERTa [2] have significantly advanced sarcasm detection by learning contextual token representations. Current research enhances transformer efficacy through the utilization of multi-head self-attention and condensed conversational context to more effectively capture pragmatic incongruity [7]. Context-aware approaches also enrich inputs with metadata or dialogue history, demonstrating that additional contextual information improves sarcasm understanding [8]. In addition to contextual modeling, the significance of tokens has been improved through TF-IDF gating and Emotion-Specific Attention (ESA), which adjust the weights of informative tokens prior to calculating attention [9]. More recently, Zhang et al. [7] conducted a comprehensive evaluation of BERT and LLM-based approaches on the SARC dataset, employing multi-headed attention to better capture context-dependent sarcasm in social media.

Sentiment Resources and Incongruity Signals: VADER is a lightweight lexicon-based sentiment analyzer designed for social media text that produces sentence- and word-level polarity scores [3]. These scores may be utilized to obtain clear sentiment incongruity characteristics that enhance transformer representations. While not adequate for sarcasm detection alone, these features offer a clear and computationally efficient indication.

Data Augmentation via Back-Translation: Back-translation creates paraphrases by translating content into a different language and back again, enhancing data variety while maintaining semantic integrity [4]. For sarcasm detection, this technique can improve generalization by reducing overfitting, although paraphrasing may occasionally weaken sarcasm-specific cues. Consequently, its effectiveness should be empirically evaluated. Furthermore, sentiment analysis and transformer-based architectures are usually examined separately instead of being integrated into a cohesive framework. Driven by these constraints, we introduce a RoBERTa architecture that is aware of incongruity, which clearly incorporates token-level sentiment incongruity characteristics into contextual representations and methodically assesses the impacts of feature injection and back-translation within a cohesive experimental framework.

3 Incongruity-Aware Sarcasm Detection Framework

The experimental pipeline follows a consistent structure for both datasets. Figure 1 provides an end-to-end view of how raw text is transformed into a trained sarcasm classifier and how performance is reported. Concretely, the pipeline proceeds as follows: (1) the dataset is loaded, cleaned, and split into train, validation and test partitions using label-stratified sampling to preserve class balance; (2) each text input is tokenized using the RoBERTa tokenizer and padded/truncated to a fixed maximum sequence length (64 tokens) to enable batching; (3) features are prepared for each example—under the baseline setting no external feature is used, while under the IDL setting a VADER-based incongruity value is computed per token and aligned to the token sequence; (4) the model is trained in PyTorch using Hugging Face Transformers with these inputs; and (5) evaluation is conducted on the held-out test split using accuracy, precision, recall, and F1-score (and AUC where applicable).

3.1 Datasets

Sarcasm cues are highly domain-dependent, we evaluate on both curated news headlines and informal Twitter text, keeping the splits label-stratified to preserve class balance and make the reported results comparable across settings.

News Headlines Sarcasm Dataset The news-headlines experiments use the *Sarcasm Headlines Dataset v2* (JSON) [10] consisting of short news headlines annotated with a binary sarcasm label. In the implementation, the headline text is stored under the `headline` field and the label under `is_sarcastic`. The dataset is split into train/validation/test using stratified sampling by label. Following the configuration in the final notebook, a subset of the dataset is used for the experiments.

Twitter Sarcasm Dataset The Twitter experiments use a tweet-level sarcasm dataset provided as CSV files (train and test) [11]. The raw columns are

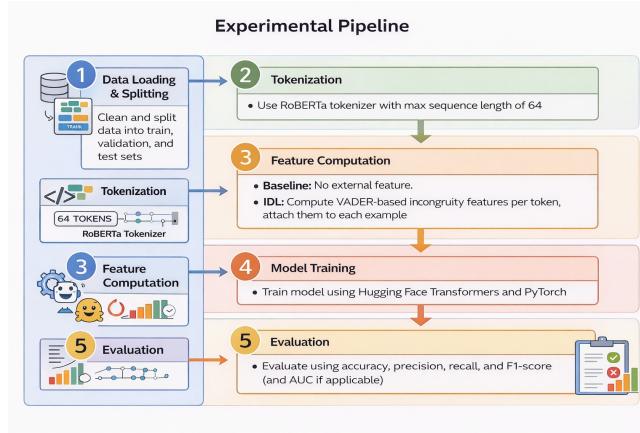


Fig. 1: Experimental pipeline used for both datasets: data splitting, RoBERTa tokenization, optional VADER-based incongruity feature computation (IDL), model training, and evaluation.

mapped to `Tweet` (text) and `is_sarcastic` (label). The training split is further divided into train/eval using a stratified split. Duplicate and empty tweets are removed during preprocessing.

3.2 Text Preprocessing

Preprocessing is intentionally minimal to preserve sarcasm cues while ensuring the data is clean enough for consistent batching and training. First, empty strings and missing values are removed so that every training instance contains valid text. Second, duplicates (especially in the Twitter training split) are removed to reduce repeated examples that can bias learning and inflate apparent performance. Finally, the cleaned text is tokenized using the RoBERTa tokenizer and truncated/padded to a fixed maximum sequence length ($L = 64$), producing uniform-length inputs (token IDs and attention masks) required by the transformer-based models used in the later stages.

3.3 Baseline Model: RoBERTa Classifier

Transformer-based models have significantly improved contextual language understanding. However, BERT suffers from static masking and the Next Sentence Prediction objective, which limits its representation capacity. RoBERTa addresses these limitations through optimized pretraining strategies, motivating its adoption in this work.

The baseline uses `roberta-base` [2] as an encoder with a sequence classification head for binary sarcasm detection. Given input token IDs and an attention mask, the model outputs logits for two classes. The baseline is trained using cross-entropy loss.

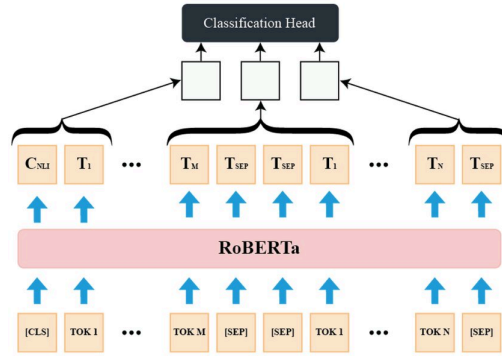


Fig. 2: RoBERTa Model

3.4 Incongruity Feature Construction (VADER)

To operationalize sentiment-based contrast, we compute token-level incongruity features using the VADER sentiment lexicon [3]. Input text is first tokenized at the word level, and each word is assigned its corresponding VADER sentiment score. These word-level scores are positionally aligned with the RoBERTa sequence and padded or truncated to length 64, providing a computationally efficient and interpretable approximation despite RoBERTa’s internal subword (BPE) tokenization. For each input text, we compute:

- Sentence-level sentiment score s using the VADER compound polarity.
- Token-level sentiment score t_i using lexicon lookup (scaled to approximately $[-1, 1]$).
- Incongruity score $f_i = |t_i - s|$.

A contrast boost is applied when strongly positive tokens appear in a negative sentence context (or vice versa), scaling the incongruity value by a factor of 1.5 under threshold conditions. The 1.5 factor does not come from a fixed theoretical rule; it was chosen as a moderate empirical boost because VADER scores are bounded in $[-1, 1]$, so smaller values would not highlight strong polarity conflict enough, while larger values could over-emphasize the lexicon signal compared with RoBERTa context. The 0.8/0.2 loss split follows standard multi-task learning practice, where the auxiliary loss is kept smaller so it regularizes the model without interfering with the main classification task [15].

Feature Alignment Strategies The external incongruity signal is represented as a fixed-length tensor aligned to the maximum sequence length $L = 64$. Token-level incongruity values are padded or truncated to form a tensor of shape $[L, 1]$, which is provided as an auxiliary input feature alongside the RoBERTa token embeddings.

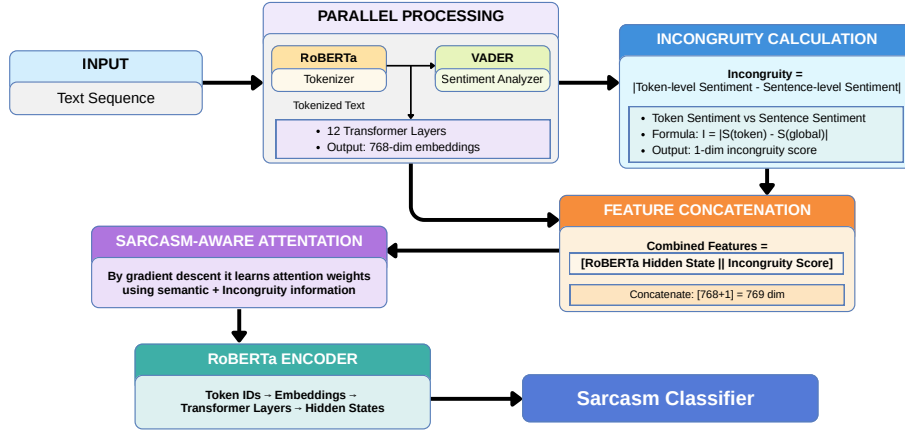


Fig. 3: IDL-RoBERTa overview: parallel RoBERTa encoding and VADER-based incongruity computation, followed by feature concatenation ($[H; F]$), sarcasm-aware attention, and context-vector formation for classification.

3.5 Proposed Model: IDL-RoBERTa

Figure 3 illustrates how the proposed IDL-RoBERTa architecture injects an explicit incongruity signal into the RoBERTa encoder pipeline. The input text sequence is processed in parallel by (i) the RoBERTa tokenizer/encoder path, which converts token IDs into contextual hidden states $H \in \mathbb{R}^{B \times L \times 768}$, and (ii) the VADER sentiment path, which produces a sentence-level sentiment score and token-level sentiment values. These sentiment values are combined into token-wise incongruity scores using $f_i = |t_i - s|$, yielding a feature tensor $F \in \mathbb{R}^{B \times L \times 1}$ aligned to the same sequence length L used by RoBERTa (with padding/truncation as needed). The model then performs feature concatenation, forming $Z = [H; F] \in \mathbb{R}^{B \times L \times 769}$ so that each token representation carries both contextual semantics and its corresponding incongruity value. Finally, a sarcasm-aware attention module learns attention weights over the augmented sequence to produce a context vector (a weighted summary of token representations), which is passed to the classifier head for sarcasm prediction.

The proposed model augments RoBERTa hidden representations with an additional incongruity feature channel. Let $H \in \mathbb{R}^{B \times L \times d}$ denote the RoBERTa hidden states ($d = 768$) and $F \in \mathbb{R}^{B \times L \times 1}$ the incongruity feature tensor. These are concatenated to form:

$$Z = [H; F] \in \mathbb{R}^{B \times L \times (d+1)}.$$

An attention layer produces a scalar score per token:

$$e_i = \mathbf{w}^\top Z_i + b,$$

followed by a softmax operation to obtain attention weights α_i . A context vector is computed as:

$$c = \sum_{i=1}^L \alpha_i H_i,$$

which is subsequently passed to a classifier head for sarcasm prediction. A complete worked example of the proposed IDL-RoBERTa has given below.

Table 1: Complete worked example of the proposed IDL-RoBERTa model. (a) RoBERTa tokenization, (b) token-level VADER sentiment scores, (c) incongruity feature computation, and (d) sarcasm-aware attention weights.

(a) RoBERTa tokenization				(b) Token-level VADER sentiment			
Position	Token	Token ID	Attention	Token	VADER raw	Normalized	Sentiment
0	<s>	0	1	Oh	0.0	0.000	Neutral
1	Oh	2285	1	great	3.1	0.775	Strong Positive
2	great	372	1	,	0.0	0.000	Neutral
3	,	6	1	another	0.0	0.000	Neutral
4	another	277	1	Monday	0.0	0.000	Neutral
5	Monday	1515	1	morning	0.0	0.000	Neutral
6	morning	662	1	meeting	0.0	0.000	Neutral
7	meeting	529	1	!	0.0	0.000	Neutral
8	!	328	1				
9	</s>	2	1				

(c) Incongruity feature computation				(d) Sarcasm-aware attention weights			
Token	t_i	s	$ t_i - s $	Token	Raw score	Softmax	Share
Oh	0.000	0.625	0.625	<s>	-0.234	0.042	4.2%
great	0.775	0.625	0.150	Oh	0.456	0.084	8.4%
,	0.000	0.625	0.625	great	0.891	0.195	19.5%
another	0.000	0.625	0.625	,	-0.123	0.047	4.7%
Monday	0.000	0.625	0.625	another	0.234	0.067	6.7%
morning	0.000	0.625	0.625	Monday	0.345	0.089	8.9%
meeting	0.000	0.625	0.625	morning	0.289	0.082	8.2%
!	0.000	0.625	0.625	meeting	0.178	0.071	7.1%
				!	0.567	0.156	15.6%
				</s>	-0.189	0.067	6.7%

Input (sarcastic): “ Oh great, another Monday morning meeting! ”

This example summarizes a full IDL-RoBERTa forward pass where RoBERTa contextual embeddings are augmented with VADER-based incongruity features.

Special tokens are added at the beginning and end. Table 1, part (a) lists the

resulting RoBERTa tokens (including special tokens) used in the worked example. Sentence-level polarity scores for the input are: positive 0.546, negative 0.000, neutral 0.454, compound 0.6249 (overall *positive* surface sentiment). Token level lexicon lookup (normalized to approximately $[-1, 1]$) is: Table 1, part (b) summarizes the token-level sentiment scores used to compute incongruity. Incongruity is computed per token as:

$$\text{Incongruity}(i) = |t_i - s|,$$

where s is the sentence compound sentiment (here $s \approx 0.625$) and t_i is the token sentiment. Table 1, part (c) shows the per-token incongruity values computed from the difference between token sentiment and sentence sentiment. The resulting feature tensor has shape $[8, 1]$ (excluding the special tokens) and mean incongruity ≈ 0.528 .

```
incongruity_features = [[0.625], [0.150], [0.625],
                        [0.625], [0.625], [0.625], [0.625], [0.625]]
mean_incongruity = 0.528
```

RoBERTa-base produces hidden states $H \in \mathbb{R}^{1 \times 10 \times 768}$ (10 tokens including special tokens). The incongruity feature (1 dimension) is concatenated per token, producing combined representations of dimension 769. An attention layer $\text{Linear}(769 \rightarrow 1)$ yields raw attention scores and softmax weights. Example values: Table 1, part (d) reports example attention scores and weights over the token sequence in the worked example. The context vector is computed as a weighted sum of token embeddings:

$$c = \sum_{i=1}^L \alpha_i H_i \in \mathbb{R}^{768}.$$

With dropout rate 0.1, the classifier $\text{Linear}(768 \rightarrow 2)$ produces logits $[0.234, 1.456]$ for [non-sarcastic, sarcastic]. Softmax gives probabilities $[0.227, 0.773]$, so the predicted label is *sarcastic* with confidence 77.3%. The auxiliary regressor predicts mean incongruity $\hat{y} = 0.512$ vs. actual $y = 0.528$ (absolute error 0.016). During training, the combined objective is:

$$\mathcal{L} = 0.8 \mathcal{L}_{CE} + 0.2 \mathcal{L}_{MSE},$$

with example values $\mathcal{L}_{CE} = 0.257$, $\mathcal{L}_{MSE} = 0.000256$, total ≈ 0.206 .

3.6 Gradient-Based Integration of External Incongruity Features

The VADER-based incongruity values are computed externally using a fixed sentiment lexicon and deterministic rules, remaining constant throughout training. Therefore, they are not learnable parameters and are not updated by gradient descent. Instead, the incongruity feature is concatenated with RoBERTa’s contextual token representations, allowing the attention layer to learn its importance. During training, gradient descent updates the attention weights based on the task loss, enabling the model to adaptively determine how strongly incongruity influences token-level attention and final sarcasm classification.

3.7 Training Configuration

All models are fine-tuned end-to-end using gradient-based optimization. In practice, training uses mini-batch gradient descent with the AdamW optimizer, updating model parameters via backpropagation to minimize the task-specific loss.

Table 2: Training hyperparameters for the News Headlines and Twitter experiments.

(a) News Headlines		(b) Twitter	
Hyperparameter Value		Hyperparameter Value	
Model	roberta-base	Model	roberta-base
Optimizer	AdamW	Optimizer	AdamW
Max length	64	Max length	64
Sample fraction	0.25	Epochs	5
Epochs	5	Batch size	16
Batch size	16	Learning rate	2×10^{-5}
Learning rate	2×10^{-5}	Warmup steps	500
Warmup steps	500	Weight decay	0.01
Weight decay	0.01	Seed	42
Seed	42	Train/eval split	0.20
Splits	Test: 0.15, Val: 0.15	BT settings	fr , batch size 8
BT settings	fr , batch size 8		

4 Experimental Results and Discussion

Models are evaluated using standard classification metrics computed on the held-out test set: accuracy, precision, recall, and F1-score. For the news-headlines setup, macro- averaged and positive-class metrics are also reported, and ROC-AUC is computed when probabilities are available. We use standard metrics for binary classification. Let TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

News Headlines Dataset: The updated News Headlines experiment notebook evaluates four RoBERTa-based configurations on the News Headlines sarcasm dataset: (i) baseline RoBERTa, (ii) IDL-RoBERTa, (iii) BT-RoBERTa, and (iv) BT+IDL-RoBERTa. The exact duplicate texts and empty entries were removed prior to the training testing split. The split was then performed with stratified sampling, so the same cleaned text could not appear in multiple partitions. For additional context, we also report CNN and LSTM baselines in the summary table. Table 5 summarizes the performance of the headline data set test-set of the baseline models and the RoBERTa-based variants.

Twitter Dataset: For Twitter, the notebook evaluates four RoBERTa-based variants: (i) baseline RoBERTa, (ii) IDL-RoBERTa, (iii) augmentation (BT)

only, and (iv) BT+IDL. For additional context, we also report CNN and LSTM baselines in the summary table. The test set contains 1,800 examples and is balanced across classes (900 sarcastic and 900 non-sarcastic), as reflected in the classification reports. Table 7 summarizes the Twitter-dataset test-set performance of the baseline models and the RoBERTa-based variants. The official test file was kept fully separate from the training file and duplicates were removed from the training side before creating the train or validation split. For back-translation experiments, augmentation was applied only to the training split ; neither validation nor test data were augmented or mixed with training data.

Table 3: Dataset versions, sources, and experimental split statistics.

Dataset	Source	Ver.	Raw	Used	Train	Val.	Test
News	Misra	v2	28,619	7,126	4,988	1,069	1,069
Headlines	& Arora [10]						
Twitter	Kaggle [11]	8361527	5,000+1,800	4,752+1,800	3,801	951	1,800

Table 4: Classification reports of the baseline RoBERTa and proposed IDL-RoBERTa models on the News Headlines dataset.

Model	Class	Precision	Recall	F1-score	Support
Baseline RoBERTa	Non-Sarcastic	0.78	0.97	0.87	574
	Sarcastic	0.96	0.69	0.80	495
	Accuracy			0.84	1069
	Macro Avg	0.87	0.83	0.84	1069
	Weighted Avg	0.86	0.84	0.84	1069
IDL-RoBERTa	Non-Sarcastic	0.91	0.91	0.91	574
	Sarcastic	0.90	0.89	0.90	495
	Accuracy			0.90	1069
	Macro Avg	0.90	0.90	0.90	1069
	Weighted Avg	0.90	0.90	0.90	1069

Figure 4 provides a compact cross-dataset view using the F1-score, contrasting the newly added CNN and LSTM baselines with the transformer baselines (RoBERTa and IDL-RoBERTa) on both News Headlines and Twitter. In general, the strongest F1 on Twitter is achieved by the combined variant BT + IDL-RoBERTa ($F1 = 0.7567$), closely followed by BT-only ($F1 = 0.7510$) and IDL-only ($F1 = 0.7342$). The baseline RoBERTa yields the highest precision (0.7536) but the lowest recall (0.5778), indicating that it misses a substantial portion of sarcastic tweets. The added CNN and LSTM baselines perform below the best transformer variants in both datasets (e.g., Twitter F1: CNN = 0.6230, LSTM = 0.6647; News Headlines F1: CNN = 0.6431, LSTM = 0.7175), reinforcing the advantage of contextual transformer encoders for sarcasm detection. In addition to the CNN, LSTM, and RoBERTa baselines already reported, we

Table 5: Test-set performance on the News Headlines Sarcasm Dataset v2.

Model	Accuracy	Recall	Precision	F1 Score
Baseline CNN	0.6688	0.6444	0.6419	0.6431
Baseline LSTM	0.7259	0.7515	0.6863	0.7175
Baseline RoBERTa	0.8419	0.6909	0.9553	0.8019
BT RoBERTa	0.8681	0.7596	0.9447	0.8421
BT+IDL-RoBERTa	0.8775	0.8141	0.9118	0.8602
BERT+MHA	0.8571	0.8276	0.8858	0.8557
IDL-RoBERTa (Proposed Model)	0.9046	0.8929	0.9002	0.8966

Table 6: Classification reports of the baseline RoBERTa and proposed IDL-RoBERTa models on the Twitter dataset.

Model	Class	Precision	Recall	F1-score	Support
Baseline RoBERTa	Non-Sarcastic	0.66	0.81	0.73	900
	Sarcastic	0.75	0.58	0.65	900
	Accuracy	—	—	0.69	1800
	Macro Avg	0.71	0.69	0.69	1800
	Weighted Avg	0.71	0.69	0.69	1800
IDL-RoBERTa	Non-Sarcastic	0.73	0.76	0.75	900
	Sarcastic	0.75	0.72	0.73	900
	Accuracy	—	—	0.74	1800
	Macro Avg	0.74	0.74	0.74	1800
	Weighted Avg	0.74	0.74	0.74	1800

Table 7: Test-set performance on the Twitter sarcasm dataset (support = 1800).

Model	Accuracy	Recall	Precision	F1 Score
Baseline CNN	0.5906	0.6767	0.5773	0.6230
Baseline LSTM	0.4978	0.9956	0.4989	0.6647
Baseline RoBERTa	0.6944	0.5778	0.7536	0.6541
BT RoBERTa	0.7344	0.8011	0.7069	0.7510
BT+IDL-RoBERTa	0.7339	0.8278	0.6969	0.7567
BERT+MHA	0.6591	0.6716	0.5539	0.6071
IDL-RoBERTa (Proposed Model)	0.7406	0.7167	0.7526	0.7342

include a comparison with the BERT + Multi-Head Attention (MHA) architecture, which represents a recent strong transformer-based baseline for sarcasm detection. Following their published configuration, including 16 attention heads, 16 transformer layers, a learning rate of 3×10^{-5} , and a sequence length of 128, we reimplemented and evaluated this model on the same datasets and splits used in our experiments. IDL-RoBERTa achieves substantially higher F1-scores

Table 8: IDL-RoBERTa multi-seed test-set summary (mean $\pm$ std, $n = 5$ seeds).

Dataset	Accuracy	Recall	Precision	F1
News Headlines	0.9044 ± 0.0089	0.8820 ± 0.0447	0.9102 ± 0.0237	0.8948 ± 0.0137
Twitter	0.7223 ± 0.0183	0.8267 ± 0.0637	0.6878 ± 0.0357	0.7484 ± 0.0049

on the same data. The VADER-based incongruity feature itself is deterministic for a fixed input, but the fine-tuning stage still depends on factors such as parameter initialization and mini-batch ordering. Therefore, we also evaluated the proposed IDL-RoBERTa model on five random seeds while keeping the data split fixed. we evaluated the proposed IDL-RoBERTa model across five random seeds (42, 123, 2024, 7, 99) on the same fixed data split. The data split itself was kept fixed across seeds; only the initialization of the model and trainer state were varied. The results are summarized in Table 8.

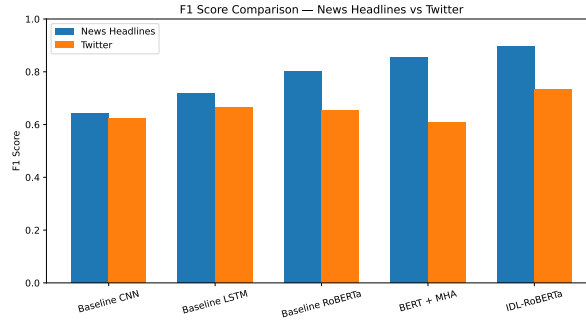


Fig. 4: F1-score comparison across the News Headlines and Twitter datasets for CNN, LSTM, RoBERTa, BERT + MHA, and IDL-RoBERTa.

Confidence Interval and Effect Size Analysis We report approximate 95% confidence intervals derived from five-seed standard deviations using $SE = \sigma/\sqrt{n}$ and a t -distribution with four degrees of freedom ($t_{0.025,4} \approx 2.123776$):

$$CI_{95\%} = \bar{x} \pm t_{0.025,4} \times \frac{\sigma}{\sqrt{n}}.$$

Both confidence intervals lie entirely above their respective baselines (Headlines: 0.8019; Twitter: 0.6541), confirming that the improvements are robust across random initialisations. Since the baseline was run with a single fixed seed, we additionally report Cohen’s d as an approximate effect size:

$$d = \frac{\bar{x}_{IDL} - x_{baseline}}{\sigma_{IDL}}.$$

For News Headlines: $d = (0.8948 - 0.8019)/0.0137 \approx \mathbf{6.78}$, and for Twitter: $d = (0.7484 - 0.6541)/0.0049 \approx \mathbf{19.24}$, both indicating very large and stable effects that provide strong statistical evidence for the reported gains.

Table 9: Approximate 95% confidence intervals for IDL-RoBERTa test F1-score across five seeds.

Dataset	Mean F1	Std	95% CI
News Headlines	0.8948	0.0137	[0.8778, 0.9118]
Twitter	0.7484	0.0049	[0.7423, 0.7545]

The experimental findings illustrate that integrating explicit sentiment incongruity into transformer-based representations reliably improves sarcasm detection accuracy in both datasets. Compared to CNN and LSTM baselines, the proposed IDL-RoBERTa achieves a stronger balance between precision and recall, resulting in superior F1 scores. Although back-translation effectively increases recall by exposing the model to more diverse linguistic expressions [12, 4], it also introduces a precision–recall trade-off due to additional false positives. The combined BT+IDL setup produces the best recall and overall F1 score, suggesting that explicit incongruity features and data enhancement work together to enhance model robustness. These results indicate that IDL-RoBERTa performs well when sarcasm is expressed through clear sentiment contrast, such as positive surface words appearing in a negative situation. In such cases, the explicit incongruity feature helps the model focus on the contrastive cue. However, the model is still challenged when sarcasm depends on external knowledge, an implicit context, or subtle pragmatic meaning. It is also less reliable when informal language or slang reduces the quality of the lexicon-based sentiment signal, especially in Twitter data. Twitter data is informal and noisy, so back-translation acts as a regularizer by adding linguistic diversity that helps the model generalize. In contrast, news headlines are concise and precisely phrased; back-translation introduces noise that can distort the original sarcastic cues and reduce the effectiveness of the incongruity feature.

5 Conclusion

Experimental findings in news headlines and Twitter datasets indicate that integrating interpretable linguistic signals into transformer representations enhances the effectiveness of sarcasm detection, while the blend of BT and IDL achieves the highest overall F1 score on Twitter. Despite these positive outcomes, multiple constraints persist. The suggested method relies on VADER, which can provide unreliable sentiment assessments for informal language and sarcasm, thus impacting the quality of the incongruity features. Future work will focus on contextual sentiment estimation, more robust augmentation techniques, and richer contextual signals to further improve incongruity-aware sarcasm detection.

Acknowledgement

This work was supported by the Center for Research, Innovation, Transformation, and Sustainability (CRITS) of Green University of Bangladesh.

Code and Data Availability

The source code is publicly available at "github.com/sayed-2299/idl-roberta-sarcasm". The experiments use publicly available news headlines and twitter sarcasm datasets; their sources and dataset preparation procedures are documented in the repository.

References

1. A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys*, vol. 50, pp. 1–22, Sept. 2017.
2. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.
3. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the 8th International AAAI Conference on Web and Social Media*, 2014.
4. S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
5. E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*, 2019.
7. L. Zhang, M. Faseeh, S. S. A. Naqvi, L. Hu, and A. Ghani, "Enhancing sarcasm detection on social media: A comprehensive study using LLMs and BERT with multi-headed attention on SARC," *PLOS ONE*, vol. 20, no. 11, p. e0334120, 2025.
8. N. A. Helal, A. Hassan, N. L. Badr, and Y. M. Afify, "A contextual-based approach for sarcasm detection," *Scientific Reports*, vol. 14, p. 15415, 2024.
9. F. Alqarni, A. Sagheer, A. Alabbad, and H. Hamdoun, "Emotion-aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition," *Scientific Reports*, vol. 15, p. 17617, 2025.
10. R. Misra and P. Arora, "Sarcasm detection using hybrid neural network," 2019.
11. Kaggle, "Twitter sarcasm dataset (train sarcasm.csv, test sarcasm.csv)." Kaggle dataset, 2026. Dataset ID: 8361527. Accessed: 2026-01-29.
12. R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.
13. Ghosh, Aniruddha, and Tony Veale. "Fracking sarcasm using neural network." *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 2016.
14. Chen, Wangqun, et al. "A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities." *Neurocomputing* 578 (2024): 127428.
15. R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.