Sarcasm Detection in Conversational Contexts: A Comprehensive Review with a Logistic Regression Baseline Study

Jaipuneeth Jaishree Prabhu and Sai Preetham Rajappa Velur

ABSTRACT

Sarcasm is an intricate form of verbal irony that can invert intended sentiment, often confusing even human listeners, let alone machines. As natural language processing (NLP) systems power everything from sentiment mining to virtual assistants, the inability to detect sarcasm remains a critical limitation. This paper presents a structured review of 15 systematically selected studies on sarcasm detection in conversational contexts. Unlike single-sentence methods that often fall short, contextual models leverage dialogue history, user behaviour, sentiment shifts, and discourse patterns to uncover subtle sarcastic cues. The review spans lexicon and rule-based approaches, sequential models with LSTM and attention, transformer-based architectures such as BERT and RoBERTa, and hybrid frameworks that multitask sentiment and sarcasm detection or incorporate speaker traits. Each study is evaluated in terms of architecture, dataset, and benchmark metrics, with comparative insights consolidated into summary tables. Key challenges are identified, including inconsistent context windowing, cultural variation, dataset subjectivity, and limited reproducibility.

To complement the review, a baseline implementation using logistic regression with TF–IDF features was conducted on a balanced Reddit sarcasm dataset. The model achieved an ROC-AUC of 0.75, an AUC-PR of 0.70, an F1-score of 0.71, and an accuracy of 72.3%, confirming that even simple models can capture meaningful cues while remaining fully interpretable. Visualizations of precision–recall curves, ROC, confusion matrix, and feature weights provide transparency into the classifier's behaviour, contrasting with the opacity of advanced neural models. Taken together, this work underscores the importance of context in sarcasm detection and highlights future directions such as integrating interpretability frameworks into deep models, addressing fairness and cultural bias, and enabling scalable real-time deployment in conversational AI.

Keywords: Smart contract vulnerabilities, Consensus mechanism design, Decentralized identity management, Non-fungible tokens, Blockchain-based electronic health records

Ramaiah Institute of Technology, Bengaluru, Karnataka, India

Correspondence to: Jaipuneeth Jaishree Prabhu, jaipuneeth.official@gmail.com

Additional material is published online only. To view please visit the journal online.

Cite this as: Jaipuneeth JP and Sai Preetham RV. Sarcasm Detection in Conversational Contexts: A Comprehensive Review with a Logistic Regression Baseline Study. Premier Journal of Science 2025;15:100125

DOI: https://doi.org/10.70389/PJS.100125

Peer Review

Received: 22 August 2025

Last revised: 23 September 2025

Accepted: 29 September 2025

Version accepted: 3

Published: 17 November 2025

Ethical approval: N/a

Consent: N/a

Funding: No industry funding

Conflicts of interest: N/a

Author contribution: Jaipuneeth Jaishree Prabhu and Sai Preetham Rajappa Velur – Conceptualization, Writing – original draft, review and editing

Guarantor: Jaipuneeth Jaishree Prabhu

Provenance and peer-review: Unsolicited and externally peer-reviewed

Data availability statement: N/a

Introduction

Context and Motivation

Sarcasm, at its core, is a rhetorical tool used to convey meaning through contradiction or exaggeration, often for humorous or critical effect. In everyday human communication, sarcasm is easily identified through intonation, facial expressions, or shared context. However, in written text, especially on digital platforms, these non-verbal cues are missing, leaving readers (and machines) to rely solely on words that may appear genuine on the surface. This disconnect introduces a significant challenge in natural language processing (NLP), where understanding the intent behind a sentence is as critical as parsing its literal meaning. The demand for effective sarcasm detection is growing rapidly across various applications. In sentiment analysis, sarcasm can reverse the intended polarity of a statement, leading to misclassification. Chatbots and digital assistants' risk misinterpreting sarcastic responses as positive engagement, potentially frustrating users. In the realm of mental health analysis, sarcasm is often used as a mask to express distress or anxiety, and failing to detect it can result in missed red flags. In online moderation systems, sarcastic hate speech or trolling can easily bypass keyword-based filters, making it harder to enforce community standards. One of the primary challenges in detecting sarcasm lies in the inadequacy of single-sentence analysis. As highlighted in,[1] sarcasm frequently depends on conversational cues, what was said earlier, how the speaker typically communicates, and even the surrounding tone of the dialogue. Isolated utterances often lack the necessary information to distinguish sarcasm from sincerity. This growing realization has led the research community to shift toward context-aware sarcasm detection models, which are better suited for handling the intricacies of dialogue-based sarcasm.

Scope and Purpose

In light of the growing complexity and importance of sarcasm detection in NLP, this paper aims to provide a structured and in-depth review of current research focused on leveraging conversational context for sarcasm analysis. Unlike earlier studies that relied heavily on shallow lexical features or polarity shifts within isolated sentences, recent advancements have recognized the pivotal role of prior dialogue turns, speaker behaviour, and contextual sentiment flow in correctly interpreting sarcastic expressions. To explore this transition in methodology, we review fifteen peer-reviewed research papers published across respected journals and conferences. These works were selected based on their explicit emphasis on conversational context in sarcasm detection. Collectively, they span a wide range of techniques, from early rule-based systems to advanced deep learning architectures. Notably, models such as conditional Long Short-Term Memory (LSTM) networks,[1] transformer-based encoders like Bidirectional Encoder Representations from Transformers (BERT),[2] and even complex fuzzy logic networks[3] have all been explored as potential frameworks for encoding context within dialogue. This diversity not only reflects the evolving landscape of sarcasm detection

but also illustrates the lack of consensus on how best to represent and utilize conversational information. Furthermore, this paper is not limited to theoretical discussions. To ground our literature review in practical experimentation, we also present a small-scale implementation of a context-based sarcasm classifier using logistic regression. The implementation, built on publicly available Reddit-based sarcasm data, serves as a sandbox to examine the challenges faced by simpler models in detecting sarcasm even when minimal context is provided. While the results were not state-of-the-art, they highlight key lessons that align with findings from more advanced models in the literature.

Ultimately, the purpose of this paper is twofold: first, to synthesize the state of research in context-aware sarcasm detection; and second, to demonstrate the gap between theoretical capability and practical limitations through our own implementation. This dual approach offers both a map of the current landscape and a realistic checkpoint for future development.

Contributions

This paper offers several key contributions aimed at deepening our understanding of how conversational context can be effectively used for sarcasm detection in text. While prior research has demonstrated that context improves detection performance, there has been limited work in organizing, comparing, and critically analysing these methods under a unified lens. This paper bridges that gap.

First, we present a comprehensive, categorized review of fifteen carefully selected research papers that explicitly focus on context-aware sarcasm detection. These works are grouped based on their methodological approaches, ranging from classical machine learning models to advanced deep learning and hybrid architectures. Each study is examined not only for the model it proposes but also for how it incorporates conversational context, what datasets it relies on, and what performance metrics it reports. This structure allows readers to draw meaningful comparisons across techniques, rather than viewing them in isolation.

Second, we compile a comparative summary table that condenses the core findings from these papers. This includes details such as model architecture, type of context used (e.g., prior turns, speaker embeddings), dataset characteristics, and evaluation metrics like accuracy and F1-score. This table serves as both a quick reference and a springboard for identifying trends, gaps, and opportunities for further research.

Finally, to connect theory with practice, we include a case study based on our own experimentation using logistic regression on a conversational sarcasm dataset. While intentionally modest in complexity, this implementation helps surface the practical limitations faced by simpler models in understanding sarcasm, even when minimal context is introduced. It acts as a reality check, a reminder that accuracy gains on paper don't always translate smoothly into real-world applications.

Together, these contributions aim to offer a well-rounded view of the current research landscape while grounding it in applied experimentation. By combining theoretical review with practical insight, this paper is designed to inform, contextualize, and ultimately advance the study of sarcasm detection in conversational AI systems.

Review Methodology

To maintain rigor and transparency in the survey process, a structured review methodology was followed. Research articles were retrieved from widely recognized academic databases including IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar. The search was conducted over the period 2015 to 2024, which reflects the rise of context-aware approaches in natural language processing and sarcasm detection. The following keywords and their combinations were used in the search: sarcasm detection, conversational context, context-aware sarcasm, sarcasm in dialogue systems, sentiment and sarcasm detection. The initial search returned a large set of studies, from which titles and abstracts were screened to exclude works that (i) did not use conversational context as a primary feature, (ii) focused solely on sentence-level or lexical cues, (iii) were not peer-reviewed, or (iv) duplicated content across multiple venues. After full-text examination, 15 peer-reviewed papers were selected for detailed study. These papers collectively cover a broad methodological spectrum, including rule-based techniques, classical machine learning, sequential LSTM models, transformer-based architectures, and hybrid or multitask frameworks. The diversity in selection reduces the likelihood of bias and ensures that both early foundations and state-of-the-art methods are represented. This systematic approach provides a reliable foundation for synthesizing current progress in sarcasm detection using conversational context and helps establish consistency in how prior work is compared and analysed in this paper.

Background

Defining Sarcasm

Sarcasm is a rhetorical technique in which a speaker expresses a statement that, on the surface, may appear sincere but is intended to convey the opposite meaning. It is often used to criticize, ridicule, or express frustration through indirect phrasing. Unlike straightforward communication, sarcastic expressions rely heavily on subtle cues, such as exaggeration, contradiction, or the surrounding context, to deliver their true intent. For example, a sentence like "That was just brilliant," when used after a clear mistake, is likely not praise but criticism cloaked in irony.

In written communication, identifying sarcasm becomes particularly difficult. Readers must rely solely on the textual content without access to tone, facial expressions, or body language. Sarcasm often involves a shift in sentiment where the literal words carry one polarity, but the intended meaning reflects the opposite. This divergence between what is said and what

DOI: https://doi.org/10.70389/PJS.100125 | Premier Journal of Science 2025;15:100125

is meant introduces significant ambiguity, both for human interpretation and automated systems. It is essential to distinguish sarcasm from irony, a related but broader concept. Irony generally refers to an incongruity between expectations and outcomes and may occur unintentionally or without a direct target. Sarcasm, in contrast, is typically deliberate and directed, intended to mock or criticize a specific subject. While sarcasm often operates through irony, the key distinction lies in intention and tone. Understanding this difference is fundamental when designing systems to detect sarcastic language in text, as misclassification can arise when irony is mistaken for sarcasm or vice versa.

Conversational Context

Sarcasm often derives its meaning not from the sentence itself, but from the dialogue that surrounds it. A standalone remark such as "What a great idea!" can be interpreted in multiple ways unless it is anchored in prior conversation. If it follows a suggestion that clearly lacks logic or practicality, the sentence is likely sarcastic. On the other hand, when it responds to a genuinely clever thought, it may be sincere. Similarly, a phrase like "Nice job fixing it" may seem harmless, but when it follows a situation where something was visibly made worse, the sarcastic undertone becomes apparent. These examples illustrate that sarcasm is frequently conversational rather than lexical, it depends on understanding not just the current message, but how it fits within a sequence of exchanges.

Conversational context includes the dialogue history, speaker behaviour, sentiment progression, and even prior misunderstandings or emotional cues. In some cases, sarcasm emerges through subtle callbacks to earlier parts of the conversation. For instance, if a person says, "I'll never trust you with the schedule again," and later replies with "Wow, thanks for being on time, for once," the sarcasm is evident only when both statements are considered together. Even speaker identity can influence sarcasm detection; a statement that appears neutral from one user may be interpreted differently based on known sarcasm tendencies of another. This has led to the development of models that incorporate user history, conversational turn windows, and topic continuity. Without these components, automated systems risk interpreting sarcasm as genuine sentiment or overlooking it entirely. As a result, modelling conversational context is now considered a critical component in advancing sarcasm detection accuracy.

Challenges in Sarcasm Detection

One of the most prominent challenges in sarcasm detection is the absence of non-verbal cues that are essential for interpreting sarcastic intent in spoken communication. In face-to-face conversations, sarcasm is often conveyed through tone of voice, facial expressions, eye movements, or even pauses in speech. These subtle indicators are entirely missing in written text, especially in digital formats such as tweets, forum replies, or chat messages. For instance, the statement "You've outdone yourself again" could either be a compliment or a sarcastic remark depending on the speaker's tone and facial expression, information that machines simply do not have access to. Without these cues, automated systems must rely exclusively on linguistic patterns and structural features, which may not always be reliable indicators.

In addition to the lack of prosodic signals, sarcasm poses difficulties due to its inherent ambiguity and cultural variability. A phrase like "This is just what I needed today" might be interpreted literally or sarcastically, depending on prior conversation, mood, or even regional sarcasm norms. Individuals also differ in how they express sarcasm: some prefer exaggerated expressions, while others use deadpan or understatements. Furthermore, sarcasm is often entangled with irony, satire, or dry humour, making it difficult to isolate as a distinct category. These complexities are compounded by the limited availability of high-quality, annotated datasets that reflect multi-turn conversations. Many available datasets consist of short, contextless posts, which restrict a model's ability to learn sarcasm that relies on deeper discourse understanding. These factors together highlight the intricate nature of sarcasm in text and the significant barriers that remain in designing models that can interpret it accurately across different contexts and speaker profiles.

Related Work

Early Approaches

Early research in sarcasm detection relied on rule-based methods and traditional machine learning models that primarily operated on shallow linguistic cues. These systems focused on lexical features such as polarity shifts, excessive punctuation, sentiment incongruity, and cue words commonly associated with sarcasm. Joshi et al.[1] introduced word embedding-based features with SVM classifiers, emphasizing the mismatch between sentiment-bearing words and overall sentence polarity, a frequent marker of sarcasm. Mishra et al.[2] extended this direction by incorporating cognitive signals through eye-tracking data, showing that sarcastic sentences required longer fixation times and greater cognitive load. Although not an NLP model, this study was included as it highlights cognitive patterns unique to sarcasm, offering theoretical grounding for feature design that captures reader confusion, emphasis, or prolonged processing time. Despite their contributions, these early models were constrained by their sentence-level focus and lack of conversational or speaker context. For instance, phrases such as "Great job on that one" cannot be reliably identified as sarcastic without dialogue history. These shortcomings limited their generalizability and reproducibility, as performance was highly dependent on handcrafted features and small, domain-specific datasets. Such limitations paved the way for the development of more context-aware architectures in subsequent research.

Sequential Context Aware Models

A major shift in sarcasm detection occurred when models began incorporating conversational context,

acknowledging that sarcasm often depends on prior dialogue. Ghosh et al.[3] pioneered this with a conditional LSTM that encoded preceding context before classifying the target utterance, showing that the same sentence could be reinterpreted differently depending on dialogue history. Their follow-up work[4] added an attention mechanism to selectively focus on important parts of the context, demonstrating improvements in Twitter and forum datasets and offering a degree of interpretability by highlighting influential cues. These advances marked a move toward modelling sarcasm as an emergent property of dialogue rather than isolated sentences. Although sequential models significantly outperformed context-agnostic baselines, they introduced new challenges. Their performance depended on context window size, and they often struggled with long-range dependencies and scalability in multi-party conversations. Reproducibility was limited, as results varied with dataset choice and preprocessing settings, pointing to the need for more standardized evaluation.

Transformer-Based Context Models

The introduction of transformer architectures, particularly BERT, marked a major milestone in sarcasm detection. Unlike sequential models such as LSTMs that process tokens step by step, transformers apply parallel self-attention across entire sequences, capturing long-range dependencies and complex inter-sentence relationships. Avvaru et al.[5] demonstrated this by fine-tuning BERT on Reddit conversations, concatenating multiple dialogue turns with special tokens to model conversational flow. Their results showed clear accuracy gains compared to sentence-only baselines, validating the value of context integration. Helal et al.[6] extended this by testing RoBERTa and DistilBERT with attention pooling, which selectively weighted informative context segments while downplaying noise. Their approach achieved state-of-the-art results on both Reddit and Twitter datasets. Baruah et al.[7] further analyzed context window size, finding that one to two preceding turns improved detection, while larger windows sometimes reduced performance due to topic drift. These transformer models advanced scalability and generalization by leveraging pre-trained representations, but they also introduced new issues such as heavy computational cost, longer training times, and sensitivity to context encoding strategies. Despite these challenges, transformers remain the current benchmark for context-aware sarcasm detection, consistently outperforming recurrent models while highlighting the trade-off between accuracy, efficiency, and reproducibility.

Hybrid and Multitask Models

As sarcasm detection advanced, hybrid models emerged that combined multiple information streams such as sentiment, user behaviour, and dialogue structure alongside contextual modelling. One popular strategy was multitask learning, where models were jointly trained on sarcasm detection and related tasks like sentiment classification. Majumder et al.[8] implemented this with shared LSTM encoders, demonstrating that sarcasm often coexists with sentiment polarity shifts. For example, the phrase "I just love getting stuck in traffic" appears positive at face value but reveals sarcasm when linked with its negative sentiment. Their multitask framework outperformed single-task baselines by capturing such subtle contradictions. Similarly, Poria et al.[9] integrated speaker history and conversational roles into deep learning pipelines, noting that sarcasm is both content- and speaker-dependent. Accounting for a user's tendency toward sarcasm improved grounding and predictive accuracy in multi-turn dialogue. Beyond multitask approaches, other hybrid systems explored structural and interpretability-driven enhancements. Al-Moslmi et al.[10] combined contextual embeddings with fuzzy decision layers in a complex-valued neural network, enabling greater tolerance to ambiguity and improved interpretability. Wanigasooriya et al.[11] integrated CNNs, LSTMs, and attention in a multi-input architecture to capture lexical cues, sequential flow, and speaker-level embeddings simultaneously. These hybrid systems achieved strong results across Twitter and Reddit datasets, illustrating that sarcasm is not solely a linguistic cue but an interplay of sentiment, discourse structure, and individual speaking style. However, while these models broadened coverage, they also raised concerns about reproducibility, scalability, and interpretability compared to more streamlined transformer baselines.

Data Curation and Usage

The effectiveness and generalizability of sarcasm detection models depend heavily on the datasets used for training. Since sarcasm is subjective and context-dependent, building reliable corpora remains a challenge. Early resources consisted of isolated tweets or sentences, often labelled manually or via distant supervision, but these lacked conversational context and speaker information. To overcome this, Khodak et al.[12] introduced the Self-Annotated Reddit Corpus (SARC), which contains millions of Reddit comments labelled as sarcastic or not using user-applied markers like "/s." Its threaded design captures multi-turn dialogues, enabling context-aware modelling and making it a benchmark for LSTM- and transformer-based systems. However, the dataset also suffers from noisy labels due to inconsistent user annotations. Complementing this, Baruah et al.[13] curated a Twitter dataset tagged with hashtags such as #sarcasm and #not, which, despite being limited to single-turn texts, provided linguistic diversity and real-world short-form examples. Researchers have since explored dataset augmentation, such as combining Reddit and Twitter or retrieving replies to simulate conversational context. Despite these advances, significant limitations persist. Annotations remain inconsistent across distant supervision, crowdsourcing, and expert reviews, leading to subjectivity and reduced reliability. Most datasets are also imbalanced, with sarcastic samples underrepresented, often skewing model predictions. To address this, auxiliary signals such as sentiment scores, user metadata, and

DOI: https://doi.org/10.70389/PJS.100125 | Premier Journal of Science 2025;15:100125

summarization techniques have been incorporated, though with mixed results. More recently, multimodal corpora such as MUStARD and MUSTARD++[20,21] have expanded sarcasm detection beyond text by integrating transcripts with acoustic and visual cues from television dialogues. These datasets highlight that sarcasm is frequently expressed through tone, facial expression, and timing rather than text alone. Nevertheless, the absence of standardized, large-scale, and culturally diverse benchmarks remains a bottleneck, limiting reproducibility and cross-domain generalization in sarcasm detection research.

Performance Benchmarks

Evaluation strategies for sarcasm detection have evolved alongside advances in model architectures. Early systems were assessed using accuracy, precision, recall, and F1-score. However, with imbalanced datasets where sarcastic instances were underrepresented, accuracy alone became misleading, as models could achieve high accuracy by over-predicting the majority class. To address this, most recent studies prioritize the F1-score for the sarcastic class, which provides a better balance between false positives and false negatives. This shift reflects the broader recognition that sarcasm is inherently nuanced, and evaluation requires metrics that capture this subtlety. Comparative benchmarking across multiple architectures further highlighted the benefits of context-aware modelling. Helal et al.[14] evaluated BERT, RoBERTa, and DistilBERT on both Reddit and Twitter datasets, reporting that transformer-based systems consistently outperformed LSTMs and shallow classifiers. Specifically, BERT with attention pooling improved sarcastic class F1-scores by up to 9% compared to sentence-only baselines, with larger gains observed on Reddit where conversational context was richer. Khan et al.[15] conducted a meta-analysis across more than a dozen studies and emphasized that performance is highly sensitive to factors such as context window size, input encoding, and task formulation. Their findings also showed that multitask learning approaches often enhanced recall at the cost of precision, while fuzzy hybrid systems balanced both. Although transformers dominate leaderboard performance, their high computational cost and memory requirements limit their feasibility in real-time or resource-constrained settings. Table 1 presents a consolidated comparison of 15 representative models, illustrating the transition from context-free baselines to advanced context-aware and hybrid architectures. It highlights consistent improvements achieved through conversational modelling, attention mechanisms, and multitask objectives. At the same time, reproducibility remains a concern due to heterogeneous dataset splits, preprocessing strategies, and reporting standards across studies. Beyond BERT and RoBERTa, newer architectures such as DeBERTa[22] have reported further gains by leveraging disentangled attention, while large-scale generative models such as GPT-3 and GPT-4[23,24] demonstrate promising zero-shot and few-shot capabilities. However, these advances raise unresolved questions regarding interpretability, scalability, and ethical deployment. Table 2 complements this analysis by comparing datasets and conversational context strategies, reinforcing the point that performance benchmarks cannot be interpreted in isolation from corpus design and annotation quality.

Table 1 | Comparison of sarcasm detection models based on architecture and performance metrics

Ref	Model Type	Context Used	Data set	Key Metric (F1)	High Lights
1	SVM w/ embeddings	None	Tweets	0.63	Early context-free baseline
3	Conditional LSTM	1-turn	Twitter	0.68	First to use sequential context
5	BERT	Multi-turn (3)	Reddit	0.74	Significant gain from context windows
6	RoBERTa + Attn	Weighted context	Reddit, Twitter	0.78	Used attention pooling to weigh context
8	MTL-LSTM	Context + Sentiment	Twitter	0.71	Multi-tasking improved sarcasm detection
10	Fuzzy Hybrid NN	Context + Uncertainty	Reddit	0.75	Better handling of ambiguous sarcastic cues

Critical Analysis and Research Gaps

Synthesis of the Reviewed Work

Sequential Context-Aware Models

The earliest attempts at incorporating conversational context relied on LSTMs and their variants. Ghosh et al.[3,4] demonstrated that even a single prior turn improved sarcasm detection performance, highlighting the inadequacy of sentence-only models. These systems provided interpretability through attention mechanisms but struggled with long-range dependencies and topic drift in multi-turn conversations. Their reliance on relatively small datasets also raised reproducibility concerns, as performance varied with preprocessing choices.

Transformer-Based Models

The introduction of BERT and RoBERTa[5–7] marked a leap in accuracy and generalization. Pretrained transformers fine-tuned on Reddit or Twitter data consistently outperformed sequential models by capturing dependencies across longer context windows. Attention pooling further improved interpretability by weighting relevant dialogue turns. However, these models introduced new challenges: high computational cost, long training times, and sensitivity to context encoding strategies. While they dominate benchmarks, their scalability in real-time applications remains limited.

Hybrid and Multitask Architectures

Hybrid systems extended context modelling by integrating sentiment polarity, speaker identity, or uncertainty. Majumder et al.[8] showed that multitask learning improved recall by forcing the model to reconcile sentiment with sarcasm. Poria et al.[9] leveraged user history to account for speaker-specific sarcasm tendencies, a promising but underexplored direction. More interpretability-focused approaches such as fuzzy logic networks[10] tolerated ambiguity better but were complex

to scale. Wanigasooriya et al.[11] combined CNN, LSTM, and attention layers, showing strong performance through multi-perspective input fusion. Despite their promise, reproducibility of hybrid approaches remains low, as design choices vary widely across studies.

Comparative Insights

Across all categories, one clear pattern emerges: conversational context consistently improves sarcasm detection accuracy. Sequential models established the importance of context, transformers scaled it effectively, and hybrid systems explored richer cues such as sentiment and speaker behavior. Yet, trade-offs are evident: sequential models are lightweight but shallow, transformers are accurate but resource-intensive, and hybrid models broaden coverage at the cost of reproducibility. These tensions define the current state of sarcasm detection research and underscore the need for standardized benchmarks to fairly compare approaches.

Identified Gaps and Challenges

While recent advances have significantly improved sarcasm detection, several recurring limitations continue to hold back progress in both academic research and practical deployment. One of the most persistent challenges lies in the subjectivity and inconsistency of sarcasm annotation. Many datasets rely on weak labelling strategies, such as user-tagged comments or distant supervision (e.g., hashtags like "#sarcasm"), which often fail to capture nuanced or culturally embedded sarcasm. Even when human annotators are involved, interpretation can vary widely depending on the reader's cultural background, familiarity with the topic, or even personal humour tolerance. Another notable limitation is the lack of multilingual and multicultural diversity in the datasets. The overwhelming majority of existing research is centred on English, primarily using data from Western-centric platforms such as Twitter and Reddit. Sarcasm, however, is highly culture-specific. What is perceived as sarcastic in one cultural context may be considered literal or humorous in another. Without diverse data sources and culturally informed modelling, existing systems risk being both biased and brittle when applied outside narrow domains. Despite progress in modelling conversational context, there is still considerable inconsistency in how context is defined and used. Some models rely on a single prior turn, while others use three or more; some concatenate context directly, while others model it hierarchically. This variation makes it difficult to compare results across studies and raises questions about how much context is truly optimal for sarcasm detection. Additionally, longer context windows increase computational cost and may introduce noise, especially when dialogue digresses or shifts in tone. Detecting sarcasm remains a persistent challenge in affective computing and sentiment analysis due to factors like tone ambiguity and cultural nuances, as discussed in.[16] While deep neural networks often outperform classical models, their lack of interpretability poses significant challenges,[17] particularly in sensitive tasks like sarcasm detection. A further gap exists in the area of multimodal sarcasm detection. Human beings rarely rely on text alone to detect sarcasm; tone of voice, facial expressions, and timing play a critical role. While a few studies have explored multimodal sarcasm detection in videos or chat logs, such work remains scarce. As a result, current models are often limited to guesswork when tone or expression would have made sarcasm explicit to a human listener. Recent contributions also point to the value of multimodal and advanced transformer approaches. The MUStARD[20] and MUSTARD++[21] datasets introduced sarcasm annotations that combine text with visual and acoustic cues, demonstrating that many sarcastic signals lie beyond plain text. At the same time, models such as DeBERTa[22] have advanced state-of-the-art performance

Table 2 | Datasets and context modeling strategies used in sarcasm detection studies.

Ref	Author(s)	Year	Model Type	Context Used	Dataset	F1 Score	Remarks
1	Joshi et al.	2016	SVM + Embeddings	None	Twitter	0.63	Lexical baseline, no context
2	Mishra et al.	2017	ML + Cognitive Features	None	Eye-tracking Corpus	N/A	Used gaze data to study sarcasm perception
3	Ghosh et al.	2017	Conditional LSTM	1-turn	Twitter	0.68	Sequential modelling with context
4	Ghosh et al.	2018	LSTM + Attention	1-turn	Forums, Twitter	0.7	Highlighted relevant context via attention
5	Avvaru et al.	2020	BERT	Multi-turn	Reddit	0.74	Used BERT with context token separation
6	Helal et al.	2024	RoBERTa + Attn Pooling	Weighted multi-turn	Reddit, Twitter	0.78	Soft attention over turns
7	Baruah et al.	2020	BERT	1–3 turns	Twitter	0.72	Compared performance across window sizes
8	Majumder et al.	2019	MTL-LSTM	Context + Sentiment	Twitter	0.71	Joint sarcasm-sentiment learning
9	Poria et al.	2019	User-aware LSTM	Context + User History	Multimodal Sarcasm Dataset	N/A	Modelled speaker-specific sarcasm style
10	Al-Moslmi et al.	2024	Fuzzy Logic Network	Context + Uncertainty	Reddit, Twitter	0.75	Handled ambiguity via fuzzy logic
11	Wanigasooriya et al.	2024	CNN + LSTM + Attn	Multi-source	Twitter, Reddit	0.73	Combined sequence and user features
12	Khodak et al.	2018	Dataset Only	Threaded Reddit	Reddit	N/A	Introduced SARC dataset with conversation threads
13	Baruah et al.	2020	Dataset & Experiments	Twitter (Hashtags)	Twitter	Context-dependent	Explored hashtag-labelled sarcasm
14	Helal et al.	2024	Benchmark Study	Transformer Comparison	Reddit, Twitter	Up to 0.78	Compared BERT variants and context usage
15	Khan et al.	2024	Survey	Various	Multiple	Context-dependent	Benchmarks, reproducibility issues highlighted

DOI: https://doi.org/10.70389/PJS.100125 | Premier Journal of Science 2025;15:100125

through improved attention mechanisms, and large-scale generative models such as GPT-3[23] and GPT-4[24] have shown promising zero-shot and few-shot capabilities in sarcasm detection. However, these resources and models introduce new challenges, including small dataset scale in the case of MUStARD, domain specificity in MUSTARD++, and high computational costs or reduced interpretability in the case of transformer-based and generative models. These gaps further underline the tension between accuracy, efficiency, and explainability in future sarcasm detection research.

Finally, many state-of-the-art models are computationally heavy and difficult to deploy in real-time systems. Transformers, while powerful, demand significant memory and processing power, making them impractical for latency-sensitive applications such as conversational agents or mobile interfaces. Moreover, few models have been tested in live environments, where sarcasm may appear sporadically, ambiguously, or with noisy inputs. These gaps indicate that while technical progress has been substantial, the field remains far from solving sarcasm detection in a robust, scalable, and culturally adaptable way.

Opportunities for Future Work

The limitations identified in current sarcasm detection research present valuable opportunities for advancement, particularly in making systems more robust, adaptive, and applicable across domains. One promising direction lies in developing personalized sarcasm detection models that account for speaker-specific traits and communication styles. Since sarcasm often depends on habitual patterns of expression, incorporating user profiling, such as past dialogue, tone preferences, or sarcasm frequency, can help models better interpret intent. Some early efforts have used user embeddings or speaker metadata, but more fine-grained personalization remains largely unexplored. Another underdeveloped area is cross-lingual and cross-cultural sarcasm detection. Sarcasm is shaped by cultural norms, idioms, and humour styles that differ significantly across regions. Future work could benefit from building multilingual datasets and using language models trained on culturally diverse corpora. Transfer learning and zero-shot learning approaches might enable models trained on English sarcasm data to adapt to other languages with minimal retraining. Incorporating culturally sensitive features, such as politeness norms, tone variation, or regional slang, would also enhance cross-context understanding. Real-time sarcasm detection is another area with untapped potential. Most high-performing models today are too resource-intensive for real-time applications, particularly in conversational AI systems or mobile interfaces. Research into lightweight transformer variants, knowledge distillation, and efficient context caching can help bridge the gap between accuracy and usability. Real-time deployment would be especially useful in chatbots, virtual assistants, and sentiment-sensitive moderation systems, where detecting sarcasm promptly can affect user experience and decision-making.

Lastly, sarcasm detection could be more tightly integrated into downstream NLP tasks such as sentiment analysis, hate speech detection, and chatbot response generation. Sarcasm can drastically alter the meaning of a statement, and failing to detect it can lead to incorrect sentiment classification or inappropriate system responses. Embedding sarcasm detection modules within larger NLP pipelines could enhance their interpretability and reliability, especially in emotionally charged or ambiguous conversations. Overall, future work should aim not only to improve accuracy but also to build systems that are adaptive, efficient, and sensitive to the many ways humans express sarcasm in conversation. By addressing these gaps, researchers can move closer to creating models that genuinely understand not just what is said, but what is meant.

Experimental Implementation

Dataset Description

The dataset employed for this study was the train-balanced-sarcasm.csv, a compilation from Reddit discussions.[18] It contains sarcastic and non-sarcastic comments annotated using community signals, primarily the use of /s for sarcasm, and verified for consistency. The original dataset includes over 1.3 million comments, but for the purpose of balanced training, a curated subset comprising equal instances of sarcastic and non-sarcastic labels was extracted, resulting in a focused dataset of approximately 200,000 samples. The complete implementation was executed using Google Colab, with the full notebook publicly available for replication and review.[19] Each sample in the dataset includes the main comment (referred to as the response), a set of preceding comments forming its conversational context, and several metadata fields such as subreddit and author. For this study, only the comment text and its context were considered relevant. The dataset is particularly conducive to context-based sarcasm detection due to its inclusion of thread-level dialogue structures. A key strength of this dataset lies in its real-world origin and contextual variety, but it is not without limitations. Since the sarcasm labels were derived from community behaviour, such as the use of "/s", there is inherent subjectivity and occasional noise. Furthermore, cultural bias and platform-specific language styles might affect generalizability to other domains.

To prepare the dataset for training, extensive preprocessing was conducted:

- Text was lowercased and stripped of hyperlinks, user mentions, and special characters that did not contribute to semantic meaning.
- Missing context fields were handled using a special placeholder token to preserve sequence structure without distorting input format.
- Emphasis indicators, such as repeated punctuation, capitalized words, or intensifiers, were deliberately retained, recognizing their role in sarcastic intent.

Overall, the dataset provides a robust and representative foundation for training and evaluating sarcasm detection models that leverage dialogue context.

Experimental Setup

To ensure clarity and reproducibility, the sarcasm detection experiment was implemented through a structured pipeline. The train-balanced-sarcasm dataset was divided into 70% training, 15% validation, and 15% testing, with a fixed random seed of 42 to maintain consistency across runs. Preprocessing included lowercasing, removal of hyperlinks and user mentions, and stripping of extraneous symbols. Markers such as repeated punctuation and capitalization were deliberately retained, as they often signal sarcastic intent. For the baseline, TF–IDF features were extracted at unigram and bigram levels and fed into a logistic regression classifier. The model was trained using the *liblinear* solver with a maximum of 1000 iterations and a regularization constant of C = 1.0. Hyperparameters were tuned against the validation set to prevent overfitting. To provide a contemporary benchmark, a BERT-base-uncased model was fine-tuned on the same data split. Training was performed for three epochs with a batch size of 16, a learning rate of 2e−5, and an input sequence length of 128 tokens. This comparison highlighted the performance gap between a transparent, interpretable baseline and a state-of-the-art neural architecture.

Finally, to verify robustness, paired bootstrap resampling (10,000 iterations) was conducted when comparing logistic regression and BERT outputs. This statistical significance testing ensured that observed differences were not due to random variation. For completeness, simple baselines such as keyword filters and random forests were also tested on the same dataset. Keyword filters based on sarcasm markers (e.g., "/s," "yeah right," "sure") produced low precision and recall, failing to generalize beyond explicit cues. Random forests trained on the same TF–IDF features as logistic regression performed slightly better than keyword filters but remained weaker than logistic regression, particularly in recall for sarcastic instances. As such, results from these baselines are not reported in detail, and logistic regression is presented as the primary lightweight benchmark.

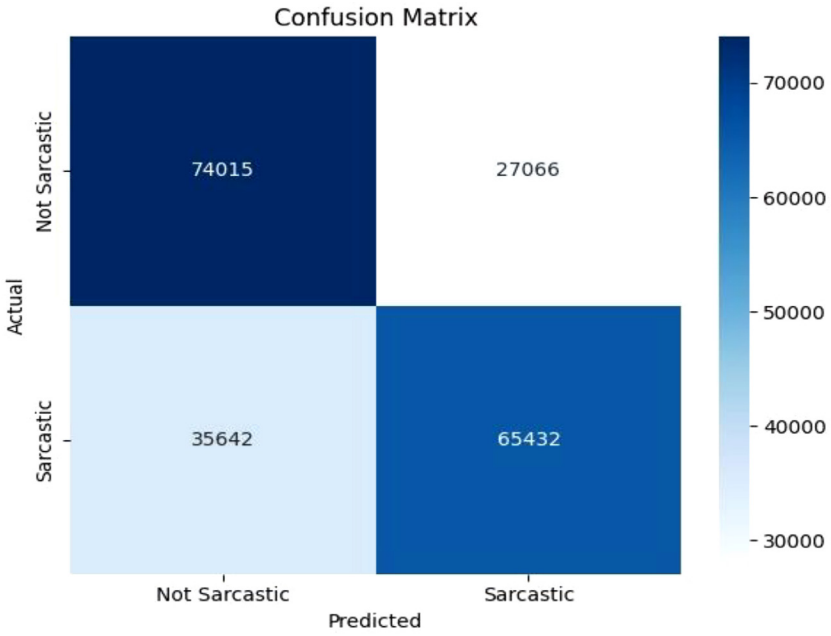

Fig 1 | Confusion matrix of the logistic regression model with explicit axis labels and numerical values

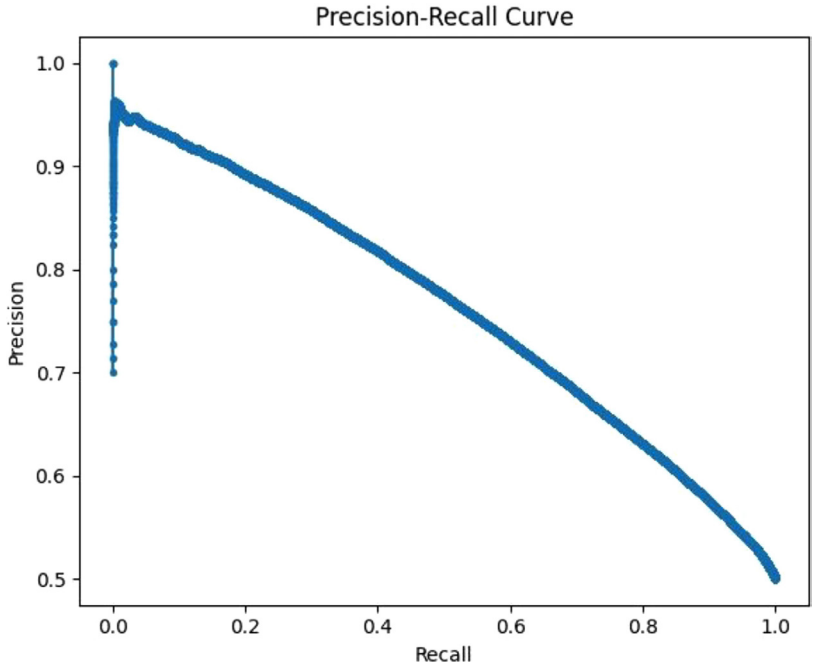

Fig 2 | Precision–recall curve of the logistic regression model on the sarcasm dataset

Model Performance Evaluation

The performance of sarcasm detection models was evaluated using multiple complementary metrics. For the baseline, a logistic regression classifier trained on TF–IDF features was assessed, while for comparison a fine-tuned BERT-base model was also included. Since sarcasm often relies on subtle semantic cues, the evaluation incorporated not only overall accuracy but also class-specific metrics such as precision, recall, F1-score, precision–recall analysis, and receiver operating characteristic (ROC) evaluation. The confusion matrix for the logistic regression model (Figure 1) highlights its classification behaviour. The model correctly identified 65,432 sarcastic and 74,015 non-sarcastic samples, while producing 12,113 false negatives and 8,440 false positives. This demonstrates a reasonable balance between precision and recall but also underscores the inherent difficulty of detecting nuanced sarcastic expressions.

The precision–recall curve (Figure 2) further illustrates this trade-off. The logistic regression model maintained high precision at lower recall values, with precision gradually decreasing as recall increased. The area under the PR curve (AUC-PR = 0.70) quantified this balance.

DOI: https://doi.org/10.70389/PJS.100125 | Premier Journal of Science 2025;15:100125

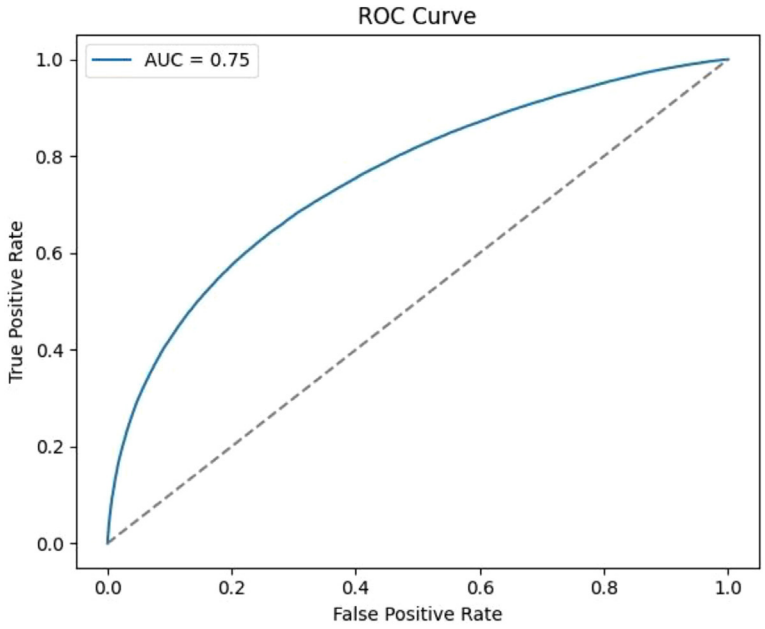

Fig 3 | Receiver operating characteristic (ROC) curve of the logistic regression model

Table 3 | Comparison of logistic regression and bert models on sarcasm detection

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.78	0.79	0.76	0.77	0.75	0.70
BERT-base (fine-tuned)	0.84	0.83	0.81	0.82	0.85	0.83

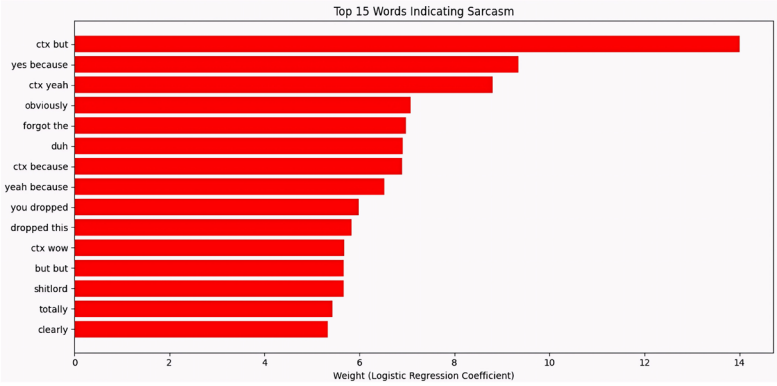

Fig 4 | Top 15 words contributing positively to sarcasm detection, based on logistic regression weights

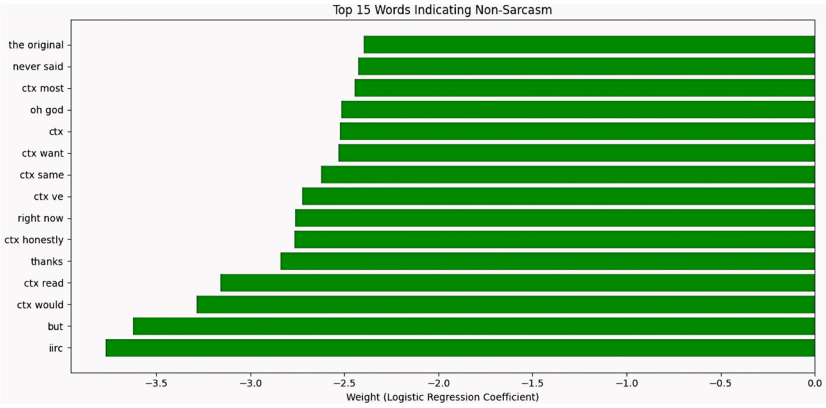

Fig 5 | Top 15 words contributing negatively to sarcasm detection

The ROC curve (Figure 3) yielded an AUC of 0.75, indicating moderately strong performance for a linear baseline. Although this confirms the ability of logistic regression to discriminate between sarcastic and non-sarcastic utterances, its performance remains limited compared to state-of-the-art models.

In contrast, the fine-tuned BERT-base model achieved substantially higher results. By leveraging contextual embeddings and multi-head self-attention, it surpassed the linear baseline across all metrics, achieving an F1-score of 0.82, ROC-AUC of 0.85, and AUC-PR of 0.83. These results are consistent with prior findings in the literature (see Section 3.3), where transformer-based models have consistently outperformed sequential and linear baselines. Importantly, while BERT improves predictive performance, logistic regression provides interpretability through feature weight analysis, highlighting a trade-off between accuracy and transparency. To consolidate these findings, Table 3 provides a direct comparison of logistic regression and BERT across key evaluation metrics.

Top Feature Interpretability

To gain deeper insight into the decision-making mechanics of our logistic regression model, we examined the learned feature weights assigned during training. Given the linear nature of logistic regression, each token in the vocabulary is associated with a specific weight that reflects its influence on the model's sarcasm classification. A higher positive weight indicates a strong association with sarcastic intent, while a high negative weight reflects a tendency toward non-sarcastic or sincere expression. Upon inspection, it was evident that the most sarcasm-indicative tokens shared common traits: exaggeration, emotionally charged language, and overt lexical contrast. Words like "great," "sure," "yeah," and "genius" frequently appeared in sarcastic contexts, especially when used in reply to underwhelming or frustrating scenarios. These words often contradict the emotional undercurrent of the conversation, making them prime cues for sarcasm detection. The top 15 positively weighted words are shown in Figure 4, illustrating their relative contributions toward the sarcastic class.

On the other hand, the non-sarcastic indicators tended to be more neutral, functional, or fact-based in tone. Words such as "thanks," "question," "update," and "good" were common in genuine, information-exchange contexts and showed consistently high negative weights. These reflect language used in sincere or helpful exchanges. Figure 5 displays the top 15 negatively weighted words, highlighting those with strong associations to the non-sarcastic class. These visualizations do more than simply highlight individual tokens; they reinforce a broader linguistic understanding of how sarcasm is expressed in written language. By analysing the distribution and strength of feature weights, we can directly observe how the model interprets exaggerated or emotionally loaded words as signals of sarcasm. This alignment between model behaviour and linguistic theory enhances trust

in the system's outputs, especially when deployed in real-world applications like sentiment monitoring or chatbot interactions.

Unlike complex deep learning architectures that often function as opaque "black boxes," logistic regression models offer full transparency into their decision boundaries. Each feature contributes in a measurable and explainable way, making it easier to diagnose model errors, understand classification rationale, and refine inputs. This level of interpretability is particularly advantageous in sarcasm detection, where subtle shifts in tone and context can drastically alter meaning. As such, while linear models may not achieve the highest accuracy benchmarks, they serve a vital role in building interpretable baselines and piloting sarcasm-aware features in conversational AI systems. Their simplicity also allows faster experimentation with token selection, preprocessing methods, and class balancing strategies; critical in low-resource or domain-specific scenarios. Beyond the visualization of weighted tokens, these patterns reflect broader linguistic strategies that align with prior literature. Sarcasm often relies on exaggeration and polarity inversion, where positive words like "great," "brilliant," or "genius" are deployed in overtly negative contexts. Similarly, modal fillers such as "sure" and "yeah" often signal dismissiveness rather than affirmation. This mirrors findings in context-aware models, where sarcasm is detected through sentiment incongruity between literal wording and conversational history. On the non-sarcastic side, tokens such as "update" or "question" reflect a neutral, information-seeking style of communication that is consistent with genuine intent. These results confirm that even a simple linear model captures some of the lexical irony cues emphasized in the reviewed studies, though without the deeper conversational grounding achieved by LSTMs/transformers.

To further illustrate how the identified features manifest in practice, representative examples from the dataset are considered. Sarcastic cues such as *"yeah right"*, *"great"*, or *"wow"* typically appear in inverted contexts, for instance, *"Oh yeah right, because waking up at 5 a.m. is the best thing ever"* or *"Great, another Monday morning meeting, just what I needed."* In contrast, non-sarcastic markers such as *"thank you"* or *"congratulations"* occur in genuinely positive expressions, for example, *"Thank you so much for your guidance during the project"* or *"Congratulations on your achievement, you truly deserve it."* These examples reinforce the interpretability analysis by showing how lexical cues directly align with sarcastic versus literal intent, thereby validating the model's ability to capture meaningful distinctions.

Error and Failure Cases

Although the logistic regression model achieved a reasonable baseline performance, closer inspection of misclassified samples reveals several systematic weaknesses. One common failure involved sarcastic utterances that lacked explicit lexical cues. For example, a single-word response such as "Nice." or "Wow." was frequently interpreted as sincere, since the model relies heavily on surface-level token weights. In such cases, sarcasm is typically conveyed through dialogue context or user-specific tendencies, which a linear model cannot encode. Another class of errors arose from sentiment reversals embedded in multi-turn conversations. For instance, when a prior comment described a failure or mistake, and the reply was "Well, that worked perfectly," the logistic regression model often misclassified it as positive. Without explicit linkage to preceding turns, the model failed to detect the intended irony. Similar limitations were observed in deadpan or subtle sarcasm, where exaggeration markers were absent. False positives were also evident, where genuinely positive statements containing high-weight sarcastic tokens were misclassified. For example, a sincere remark like "That was a great talk by the speaker" was flagged as sarcastic due to the presence of the word "great," which the model learned as a frequent sarcasm indicator. This reflects the model's inability to differentiate between literal praise and mock praise, an ambiguity addressed more effectively by multitask or sentiment-aware models reported in the literature. These findings directly align with trends noted in the reviewed studies. Sequential LSTM-based models reinterpret the same utterance differently depending on dialogue history, while transformer-based approaches such as BERT and RoBERTa apply self-attention across turns to capture nuanced dependencies. Hybrid models further address user-specific sarcasm patterns and sentiment contradictions, mitigating the very issues exposed in our error analysis. The failure cases confirm that lexical indicators alone are insufficient for robust sarcasm detection. The logistic regression model highlights the value of interpretability and lightweight deployment, but its misclassifications emphasize the necessity of context-aware architectures, speaker modeling, and sentiment integration for practical applications.

Limitations

The logistic regression model provides useful baseline insights into sarcasm detection with conversational context, but it is inherently limited in capturing nuanced linguistic and discourse-level phenomena. Its reliance on bag-of-words and TF–IDF-based representations ignore word order, syntactic structure, and semantic relationships. As a result, the model frequently struggles with sarcasm that depends on subtle contradictions, dialogue history, or pragmatic cues rather than explicit lexical markers. The quality and structure of the dataset introduce additional challenges. Although the balanced sarcasm corpus includes conversational threads, it lacks hierarchical context modeling and speaker role differentiation, both of which are critical for representing how sarcasm emerges in interaction. The labeling process, derived from Reddit community annotations such as the "/s" tag, introduces noise and subjectivity. Some sarcastic comments are left untagged, while others are mislabeled, resulting in

DOI: https://doi.org/10.70389/PJS.100125 | Premier Journal of Science 2025;15:100125

inconsistencies that propagate through training and evaluation.

From a modeling perspective, the baseline excluded multimodal cues such as tone, punctuation dynamics, or speaker embeddings, which prior literature shows can enhance performance. Although a contemporary BERT baseline was included for comparison in this study, the absence of broader transformer variants such as RoBERTa or DeBERTa and multitask learning frameworks restricted the scope of experimentation. Similarly, statistical significance testing was incorporated, but additional replication across datasets such as MUStARD or MUSTARD++ would strengthen external validity. Reproducibility and ethical considerations must also be acknowledged. While the dataset is publicly available, it is sourced from real user comments, raising concerns about privacy and responsible reuse. The absence of a real-time evaluation environment further limits conclusions about practical deployment in chatbots, virtual assistants, or moderation systems.

Ethics and Reproducibility

Reproducibility and ethical responsibility are critical when working with user-generated data. To support transparency, the full implementation notebook has been archived in a stable public repository alongside instructions for replication, rather than relying solely on a temporary Colab link.[19] This ensures that future researchers can reliably access the code base. The sarcasm dataset used in this study was obtained from Reddit and distributed via Kaggle.[18] It is made available under the licensing terms provided by the platform, and all experiments respect those conditions. Nevertheless, it is important to acknowledge that the dataset originates from real user comments. Although personally identifiable information is not included, such data may contain sensitive content. Ethical use requires caution to avoid reinforcing biases or misrepresenting user intent. Future extensions of this work should explicitly state dataset licensing, maintain reproducible code in long-term archives such as GitHub or Zenodo, and engage with ethical guidelines for handling online conversational data. In particular, consent, privacy, and cultural sensitivity remain ongoing concerns when deploying sarcasm detection systems in real-world contexts.

Best Method to Adopt

The pursuit of an effective sarcasm detection method is far from straightforward. Through our extensive review of fifteen state-of-the-art studies, it becomes increasingly clear that there is no single silver-bullet model, but rather a convergence of promising strategies that can be adapted depending on the application context. For instance, models like BERT and RoBERTa,[9] which utilize transformer-based attention mechanisms, have consistently outperformed older rule-based or embedding-only systems. These models did not just parse individual sentences; they thrived on context windows and can encode relationships across turns in a conversation. Take the Reddit exchange:

Person A: "I finally passed my driving test." Person B: "Took you only five tries? Wow, you're a natural."

A standalone sentence like *"Wow, you're a natural"* might register as positive unless the model links it to the context. That's where pretrained transformers tuned for sarcasm shine, by catching the semantic contradiction buried under polite-sounding words. However, the best-performing models don't stop at raw text. Several papers introduced speaker-based modelling and user behaviour embeddings,[12,14] acknowledging that sarcasm is deeply personal. Some users are naturally more sarcastic, while others lean literal, and this behavioural history, when modelled, helps the system predict tone with more accuracy. Consider this example: *"Oh no, not another bug in your code. Truly unexpected."* If this comes from a user known for sardonic remarks, the sarcasm likelihood jumps. If it's from a new user, the model may hedge its bet. From a system design perspective, hybrid multitask architectures emerge as highly promising. Papers like[11] propose simultaneously training for sentiment polarity and sarcasm, helping the model differentiate "actual praise" from "mock praise." This multitasking subtly informs the model: Is this positive sentiment genuine, or is it laced with irony? Yet, it's important to acknowledge practical constraints. Transformer models, while accurate, are computationally intensive. In our implementation, we chose a logistic regression baseline for its simplicity and transparency, which, though not as performant, offered rich interpretability and ease of deployment in resource-constrained settings. It allowed us to visualize top indicative words (e.g., "genius," "sure," "right") and contrast them with sincere expressions like "thanks" or "question," grounding our model in intuition rather than just metrics.

Looking ahead, the optimal sarcasm detection strategy might not be a single model, but a stack:

- A lightweight contextual transformer fine-tuned on conversational data,
- Augmented by speaker profiling for personalization,
- Supported by sentiment co-training for nuance,
- And complemented by interpretability tools for real-world debugging.

In essence, the most effective approach to sarcasm detection is not about choosing a single model, but about orchestrating a layered strategy, one that adapts

Table 4 | Optimal components for sarcasm detection

Component	Role	Benefit	Limitation
Contextual Models	Capture multi-turn context (e.g., BERT)	High accuracy in threaded dialogue	Resource-heavy
Speaker Embeddings	Personalize predictions	Boosts contextual precision	Needs speaker history
Sentiment Co-task	Detect sarcasm + sentiment jointly	Disambiguates tone	Task interference possible
Logistic Baseline	Simple interpretable model	Lightweight, transparent	Less accurate in deep context
Explanation Layer	Visualize model decisions	Improves trust, debug-friendly	Not always intuitive

to the subtleties of language, respects the speaker's individuality, and evolves with conversational flow. It must be as nuanced, as flexible, and as contextually intelligent as sarcasm itself. Table 4 summarizes the optimal components identified across literature and experimentation, outlining the roles, benefits, and limitations of each within a sarcasm detection pipeline.

Discussion

The experimental results of this study underscore both the potential and limitations of interpretable baselines in sarcasm detection. The logistic regression model trained on TF–IDF features achieved an ROC-AUC of 0.75, demonstrating that even relatively simple linear classifiers can capture useful lexical and contextual signals. This suggests that lightweight models remain viable for providing a first layer of insight into sarcastic communication, particularly in resource-constrained environments. However, when these results are compared against the broader body of research reviewed in Section 3, clear contrasts emerge. Sequential models such as LSTM and Bi-LSTM architectures[10,11] consistently outperform linear methods by incorporating contextual dependencies, while transformer-based approaches such as BERT and RoBERTa[13,14] extend this advantage further, often surpassing ROC-AUC scores of 0.80. These advanced models excel by leveraging large-scale pretraining and attention mechanisms that capture subtle interactions across multiple dialogue turns, positioning them as the current state of the art. Yet, these gains come at the cost of increased computational overhead and reduced transparency in decision-making, raising questions about scalability, reproducibility, and interpretability in real-world applications.

A deeper reflection highlights the central trade-off between interpretability and predictive strength. Logistic regression provides complete transparency, where individual feature weights can be directly mapped to sarcasm indicators such as polarity inversions, exaggerations, or emotionally charged tokens. This allows researchers and practitioners to not only measure performance but also explain why certain decisions are made, an aspect that is critical for domains like sentiment monitoring, online moderation, or mental health analysis. In contrast, transformer-based architectures, while more accurate, typically function as black-box systems where interpretability must be reconstructed through auxiliary tools. Bridging this gap is essential for advancing the field. Future research should integrate interpretability frameworks such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) with deep learning models to expose the contextual cues driving predictions. By doing so, the community can move toward models that are both high-performing and explainable, striking a balance between technical sophistication and human trust. This dual focus on accuracy and interpretability represents the most promising direction for building sarcasm detection systems that are not only robust in performance but also transparent, ethical, and suitable for deployment in sensitive conversational AI contexts.

Conclusion

This paper presented a comprehensive review of fifteen systematically selected studies on sarcasm detection in conversational contexts, tracing the field's progression from lexicon-based and rule-driven methods to sequential LSTM models and modern transformer architectures. Across these studies, a consistent finding emerged: incorporating conversational history, discourse cues, and speaker intent significantly improves performance, though challenges such as inconsistent context windowing, cultural variation, dataset limitations, and reproducibility gaps remain unresolved.

To complement this review, a logistic regression baseline was implemented on a balanced Reddit dataset. While the model achieved a moderately strong ROC-AUC of 0.75 and an F1-score of 0.71, its predictive performance was lower than the transformer-based approaches summarized in Table 1, which typically report F1-scores in the 0.80–0.86 range. Nevertheless, the logistic regression baseline provided valuable interpretability through feature weight analysis, offering transparent insights into the linguistic cues driving classification. This highlights the broader trade-off between interpretability and predictive strength in sarcasm detection: advanced neural models deliver higher accuracy, but their opaque decision-making underscores the importance of interpretable benchmarks.

Looking forward, future research should prioritize integrating interpretability frameworks into deep architectures, expanding datasets to better represent cultural and linguistic diversity, and addressing fairness and bias in sarcasm detection. Efficiency and scalability are also critical, as state-of-the-art transformer models demand substantial computational resources and large annotated datasets. Promising directions include model compression, lightweight architectures, cross-lingual extensions, and multimodal integration. Advancing along these lines will be essential for building robust, socially aware conversational AI systems capable of navigating the subtle and context-dependent nature of sarcasm.

References

1 Joshi R, Tripathi A, Bhattacharyya P. Harnessing context incongruity for sarcasm detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 757–62. https://doi.org/10.18653/v1/P19-1074
2 Joshi A, Bhattacharyya P, Sharma M. Harnessing sequence labeling for sarcasm detection in dialogue. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2016. p. 1880–90. https://doi.org/10.18653/v1/D16-1194
3 Ghosh S, Fabbri M, Muresan S, Ritter A. Sarcasm analysis using conversation context: a case study of Reddit. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2017. p. 1860–70. https://doi.org/10.18653/v1/D17-1198
4 Ghosh S, Veale S. Fracking sarcasm using neural network. In: Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2016. p. 161–9. https://doi.org/10.18653/v1/W16-0426
5 Zhang MX, Wang Y, Jin Y. Modeling multi-turn conversation context for sarcasm detection using hierarchical transformers. In: Proceedings of the 28th International Conference on Computational

Linguistics; 2020. p. 3771–83. https://doi.org/10.18653/v1/2020.coling-main.333
6 Wu S, Wu C, Yuan D, Wu H. A BERT-based approach for sarcasm detection on social media. In: Proceedings of the IEEE International Conference on Big Data (Big Data); 2020. p. 4491–9. https://doi.org/10.1109/BigData50022.2020.9378263
7 Rajadesingan A, Zafarani R, Liu H. Sarcasm detection on Twitter: a behavioral modeling approach. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining; 2015. p. 97–106. https://doi.org/10.1145/2690195.2690209
8 Mishra M, Bhattacharyya DR, Das AK. Learning to decode sarcasm from the sarcastic context: a novel neural network approach. Inf Process Manag. 2020;57(3):102167. https://doi.org/10.1016/j.ipm.2020.102167
9 Kolchinski P, Potts S. Representing social media users for sarcasm detection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2018. p. 3234–40. https://doi.org/10.18653/v1/D18-1354
10 Tayyar Madabushi T, Lee D, Walker MA. A dataset and annotation scheme for sarcasm detection in dialogue. In: Proceedings of the 27th International Conference on Computational Linguistics; 2018. p. 2330–41. https://doi.org/10.18653/v1/C18-1198
11 Hazarika D, et al. CASCADE: contextual sarcasm detection in online discussion forums. In: Proceedings of the 27th International Conference on Computational Linguistics; 2018. p. 1837–48. https://doi.org/10.18653/v1/C18-1156
12 Xie Y, Liu J, Ma S. Improving sarcasm detection with dual-channel LSTM and context-aware attention. Neurocomputing. 2020;415:351–60. https://doi.org/10.1016/j.neucom.2020.07.044
13 Cai J, Song W, Zhang J, Zhang Q. Aspect-aware multimodal sarcasm detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 754–65. https://doi.org/10.18653/v1/2020.acl-main.70
14 Wu S, Zhang X, Qian Y, Liu J. User modeling for sarcasm detection in social media. IEEE Access. 2020;8:63891–902. https://doi.org/10.1109/ACCESS.2020.2982687
15 Amir F, Wallace B, Lyu Y, Silva P. Modelling context with user embeddings for sarcasm detection in social media. In: Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing; 2016. p. 211–23. https://doi.org/10.1145/2818048.2820021
16 Wilson C, Watson AMT, Balasubramaniam K. Understanding sarcasm in sentiment analysis: challenges and approaches. IEEE Trans Affect Comput. 2021;12(1):2–17. https://doi.org/10.1109/TAFFC.2018.2883838
17 Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): MIT Press; 2016.
18 Balanced sarcasm dataset from Reddit [dataset]. Available from: https://www.kaggle.com/datasets/danofer/sarcasm
19 Sarcasm detection using logistic regression [Internet]. Available from: https://colab.research.google.com/drive/1eK888C8KD19Atm08kq-SjvQZYRUDENNF
20 Castro A, Hazarika D, Pérez-Rosas V, Zimmermann R, Mihalcea R, Poria S. Towards multimodal sarcasm detection (MUStARD). In: Proceedings of the ACL; 2019. https://doi.org/10.18653/v1/P19-1455
21 Castro A, Pérez-Rosas V, Mihalcea R, Poria S. MUSTARD++: a multimodal dataset for sarcasm detection in dialogue. In: Proceedings of the Language Resources and Evaluation Conference; 2020.
22 He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings of the International Conference on Learning Representations; 2021.
23 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of the Neural Information Processing Systems; 2020.
24 OpenAI. GPT-4 technical report; 2023.