International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 936-958
https://doi.org/10.70917/ijcisim-2026-4493
Article

Sarcasm Detection Using Learning Techniques in a Social Engineering Context: A PRISMA-Based Systematic Review

Yuvraj G. Nikam[1], Vijay Pal Singh[2], Dhanshri Amol Shinde[3]*

[1]Department of Computer Engineering and Applications, Faculty of Engineering and Technology, Mangalayatan University, Beswan, Aligarh, Uttar Pradesh, India.
Email: 20231919_yuvrajs@mangalayatan.edu.in
[2]Department of Computer Engineering and Applications, Faculty of Engineering and Technology, Mangalayatan University, Beswan, Aligarh, Uttar Pradesh, India.
Email: vptilotiya@gmail.com
[3]Department of Computer Engineering, Tolani Maritime Institute, Pune – 410507, Maharashtra, India.
Email: patildhanshri@gmail.com

Abstract: Sarcasm—the deliberate expression of a meaning opposite to the literal content of an utterance—is pervasive in social-media communication and constitutes a major source of error for sentiment analysis, opinion mining and manipulation-detection systems deployed in social-engineering contexts. Because sarcasm inverts surface polarity, its automatic detection remains a hard, context-dependent natural-language-processing (NLP) problem. This paper presents a systematic review conducted in accordance with the PRISMA 2020 guidelines. Eight bibliographic databases were searched for the period 2010–2026; from 1,320 identified records, a four-phase identification–screening–eligibility–inclusion process yielded 40 primary studies for synthesis. Beyond the review protocol, the paper provides an extended theoretical foundation covering pragmatic theories of irony, text representation, classical machine-learning classifiers, recurrent and convolutional architectures, the attention and Transformer formalism, pre-trained language models, multimodal fusion and evaluation metrics. The synthesis is organised around four research questions addressing (RQ1) benchmark datasets and social-media text sources, (RQ2) NLP pre-processing techniques, (RQ3) machine-learning and deep-learning models, and (RQ4) transformer-based and context-aware models, with a dedicated comparison at each question and a per-study questionnaire-style extraction form. Results reveal a clear methodological progression from feature-engineered classifiers, through CNN/RNN and attention hybrids, to transformer and multimodal graph architectures, with contextual embeddings consistently improving performance. Persistent gaps include weak modelling of extra-utterance context, noisy distant-supervision labels, severe class imbalance in intended-sarcasm corpora, limited cross-domain and code-mixed generalisation, and poor interpretability. On this basis we outline a precision-aware, context-integrating sarcasm-prediction algorithm and an evaluation protocol for validation against the state of the art.

Keywords: sarcasm detection; irony detection; systematic literature review; PRISMA; natural language processing; deep learning; transformer models; BERT; social media; social engineering; sentiment analysis

1. INTRODUCTION

1.1. Background and Motivation

The growth of microblogging and social-networking platforms has made user-generated text one of the richest available sources of public opinion. Automated systems increasingly mine this text to gauge sentiment, detect abuse and identify manipulation attempts. Yet a substantial fraction of social-media language is figurative, and sarcasm in

particular systematically inverts the apparent polarity of a message. A post such as "Great, another software update that breaks everything—exactly what I needed today" carries strongly positive surface vocabulary but conveys a negative intent. Systems that read such text literally produce confident but incorrect conclusions, and the error is systematic rather than random [1,2].

In a social-engineering context this failure is more than a nuisance. Social-engineering and online-manipulation campaigns rely on tone, insinuation and deniable irony; sarcasm is a common vehicle for veiled hostility, deception and covert persuasion, and it provides plausible deniability to an attacker whose message is challenged. Reliable sarcasm detection therefore contributes directly to safer sentiment analytics, more accurate abuse- and manipulation-detection pipelines, and better inference of the intent behind social-media messages.

1.2. The Sarcasm-Detection Challenge

Sarcasm is difficult to model for several intertwined reasons. First, it is inherently context-dependent: identical wording may be sincere or sarcastic depending on author history, conversational thread and world knowledge [11,16,18]. Second, social-media text is short, noisy and saturated with slang, emojis, hashtags and code-mixing, which complicates pre-processing and representation [8,17]. Third, gold labels are difficult to obtain—distant supervision via hashtags is noisy, third-party annotation is subjective, and even author-provided "intended" labels yield highly imbalanced corpora [22,30,33]. Fourth, sarcasm is culturally and demographically variable, so models trained on one community transfer poorly to another. Together these factors make sarcasm detection an enduring benchmark for context-aware NLP.

1.3. Objectives of This Review

This systematic review supports a broader research programme whose objectives are:

1. To analyse the performance of existing learning approaches for sarcasm detection and perform a gap analysis;
2. To design a precision-aware mechanism for identifying the context of text;
3. To propose an accurate sarcasm-prediction algorithm that integrates textual context and is capable of relating it to the general problem; and
4. To evaluate the performance of the proposed algorithm and validate the findings in comparison with the state of the art.

1.4. Contributions

- An extended theoretical foundation (Section 2) covering pragmatic theories of irony, representation learning, classical and deep architectures, the Transformer formalism and evaluation metrics, so that the review is self-contained for readers entering the field.
- A reproducible PRISMA 2020 protocol with explicit search strings, database coverage, detailed inclusion/exclusion criteria with rationale, a quality-assessment rubric and a full flow diagram.
- A four-question synthesis of 40 primary studies with a dedicated comparison table and comparative discussion at each research question.
- An extended, era-structured literature review complemented by a per-study questionnaire-style extraction form enabling item-by-item comparison.
- A structured gap analysis and a proposed precision-aware, context-integrating sarcasm-prediction framework with a concrete validation plan.

1.5. Organisation of the Paper

Section 2 develops the theoretical background. Section 3 details the PRISMA review methodology, including research questions, search strategy, inclusion and exclusion criteria, the flow diagram and quality assessment. Section 4 presents results organised by research question with a comparison at each. Section 5 provides the extended literature review in narrative and questionnaire form. Section 6 discusses findings and performs the gap analysis. Section 7 outlines the proposed research direction and Section 8 concludes.

2. BACKGROUND

This section establishes the conceptual and mathematical vocabulary used throughout the review. It proceeds from the linguistic characterisation of sarcasm, through text representation and learning architectures, to evaluation.

2.1. Linguistic and Pragmatic Theories of Sarcasm

Sarcasm is a form of verbal irony in which the speaker's intended meaning is opposed to, or sharply divergent from, the literal meaning, typically with a critical or mocking edge directed at a target. Computational work draws on four principal pragmatic accounts.

2.1.1. Gricean Maxim Violation

Grice's cooperative principle [49] holds that speakers normally observe maxims of quality, quantity, relation and manner. Irony is analysed as a deliberate, ostentatious flouting of the maxim of quality: the speaker says something patently false, inviting the hearer to derive an implicature. Computationally this motivates features that detect a mismatch between an utterance and observable reality or context.

2.1.2. Echoic Mention Theory

Sperber and Wilson [50] argue that ironic utterances echo a previously expressed or attributed proposition while signalling a dissociative attitude towards it. This account explains why conversational context and author history carry so much predictive weight, and directly underpins context-aware models that encode the preceding thread [18,28].

2.1.3. Pretense Theory

Clark and Gerrig [51] propose that the ironist pretends to be an injudicious speaker addressing an uncomprehending audience, expecting the real audience to see through the pretense. This framing motivates modelling speaker persona and audience, as in user-embedding approaches [12,16].

2.1.4. Muting and Contrast

Dews and Winner [52] show that irony mutes the literal criticism it conveys, serving a social face-saving function; Gibbs [53] documents its prevalence in friendly talk. The associated computational construct is

context incongruity—the co-occurrence of positive surface sentiment with a negative situation [6,9]. Formally, given an utterance with sentiment polarity *s(u)* and a situational or contextual polarity *s(k)*, an incongruity score may be defined as in Equation (1):

$$I(u,k) = ½ \cdot |\, s(u) - s(k) \,|, \; s(\cdot) \in [-1,+1] \quad (1)$$

Larger values of I indicate stronger polarity contrast and hence stronger evidence of sarcasm. Almost every model surveyed in this review can be read as a progressively more sophisticated estimator of this quantity.

2.2. Sarcasm, Irony, Satire and Humour

These phenomena overlap but are not identical, and conflating them harms both dataset construction and evaluation. Irony is the broad category of meaning inversion; sarcasm is irony with a critical target and, usually, an identifiable victim; satire is a genre-level device that uses irony for social critique (the basis of the News Headlines dataset [24]); and humour is a superordinate class in which incongruity is resolved for amusement rather than criticism. Sub-types matter too: self-deprecating sarcasm, whose target is the speaker, requires distinct modelling and has motivated dedicated architectures [23,32].

2.3. Sarcasm in the Social-Engineering Context

Social engineering is the manipulation of people into divulging information or performing actions against their interest. Figurative language serves such manipulation in three ways relevant to this review. First, it defeats automated moderation: sarcastic hostility passes polarity filters that key on surface sentiment. Second, it provides deniability, allowing an attacker to reframe a probe or insult as a joke when challenged. Third, it signals in-group membership, which pretexting exploits to build rapport. Accurate sarcasm prediction therefore functions as an intent-inference component within a broader manipulation-detection pipeline, and this framing motivates the emphasis this review places on precision and interpretability rather than raw accuracy.

2.4. NLP Pre-Processing Theory

Pre-processing maps raw social text to model-ready units. Normalisation removes or canonicalises URLs, mentions and HTML entities. Tokenisation segments text; modern systems use sub-word schemes—Byte-Pair Encoding and WordPiece—that represent an out-of-vocabulary slang term as a sequence of known sub-word units, which is essential for noisy social text. Stemming and lemmatisation reduce inflectional variance for classical pipelines but are usually omitted for transformers, since sub-word models exploit morphology directly. Sarcasm-specific steps include hashtag segmentation (recovering words from concatenated tokens), emoji and emoticon preservation (affective cues carry polarity contrast [17]), and retention of punctuation, capitalisation and character elongation, which are pragmatic markers rather than noise [8].

2.5. Text Representation and Feature Engineering

Classical pipelines represent a document as a sparse vector. Term frequency–inverse document frequency weights a term t in document d over a corpus D as in Equation (2):

$$tf - idf(t,d,D) = tf(t,d) \cdot log(\,|D| \,/\, (1 + |\{d' \in D : t \in d'\}|)\,) \qquad (2)$$

Distributional models replace sparse counts with dense vectors. Word2Vec [54] learns embeddings by maximising the log-probability of context words given a target word, as in Equation (3):

$$J = (1/T)\, \Sigma_t\, \Sigma_j\, log\; p(w_{t+j} \mid w_t), \quad -m \le j \le m,\; j \ne 0 \qquad (3)$$

GloVe [55] instead factorises a global co-occurrence matrix. Both yield a single static vector per word type, which is a fundamental limitation for sarcasm: the word "great" receives one representation regardless of whether it is sincere or sarcastic. Contextual embeddings resolve this by conditioning the vector on the surrounding sequence, which is the principal reason they improved sarcasm detection so markedly [20,29].

2.6. Classical Machine-Learning Classifiers

Given feature vectors $x \in R^n$ and labels y ∈ {sarcastic, non-sarcastic}, early work applied standard discriminative and generative classifiers. A linear support vector machine seeks the maximum-margin hyperplane by minimising the objective in Equation (4):

$$min\ over\ (w,b):\ \tfrac{1}{2}\|w\|^2 + C\, \Sigma_i\, max(0, 1 - y_i\,(w^T x_i + b)) \qquad (4)$$

Logistic regression models the posterior directly as $\sigma(w^T x + b)$ with σ the logistic function; multinomial Naïve Bayes assumes conditional feature independence; random forests aggregate decorrelated decision trees, which suits the heterogeneous hand-crafted feature sets typical of this era [8,38]. These models are fast, interpretable and reproducible, but their performance is bounded by the quality of manual feature engineering and they cannot represent word order or long-range context.

2.7. Deep Sequence and Convolutional Architectures

Convolutional neural networks apply learned filters over an embedded sentence matrix, extracting position-invariant n-gram features. A feature map from filter w over window h is given by Equation (5):

$$c_i = f(\,w \cdot x[i : i+h-1] + b\,), \quad \hat{c} = max\{\,c_1, \ldots, c_m\,\},\; m = n - h + 1 \qquad (5)$$

Recurrent networks instead process tokens sequentially, maintaining a hidden state. Long short-term memory units [56] mitigate vanishing gradients through gated cells, as in Equation (6):

$$f_t = \sigma(Wf \cdot [h_{t-1}, x_t] + bf), \quad i_t = \sigma(Wi \cdot [h_{t-1}, x_t] + bi), \quad o_t = \sigma(Wo \cdot [h_{t-1}, x_t] + bo) \qquad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(WC \cdot [h_{t-1}, x_t] + bC), \quad h_t = o_t \odot tanh(C_t) \qquad (7)$$

Gated recurrent units simplify this to update and reset gates. Bidirectional variants concatenate forward and backward states so that each token is contextualised by both its left and right neighbours—valuable for sarcasm, where the incongruent cue may follow the sentiment-bearing span. The seminal CNN–LSTM–DNN cascade of Ghosh and Veale [13] combined both inductive biases and set the template for the deep-learning era.

2.8. Attention and the Transformer

Attention lets a model weight tokens by relevance rather than by recency. Scaled dot-product attention over queries Q, keys K and values V is defined in Equation (8) [41]:

$$Attention(Q,K,V) = softmax(\, Q\, K^{T} / \sqrt{d\_k}\,)\, V \qquad (8)$$

Multi-head attention runs h such projections in parallel and concatenates them, as in Equation (9), allowing distinct heads to specialise—empirically, some heads in sarcasm models align with the incongruent span, which is the basis of interpretable architectures [31]:

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_h)\, W^{O}, \quad head_i = Attention(Q\, W_i^{Q}, K\, W_i^{K}, V\, W_i^{V}) \qquad (9)$$

Because self-attention is order-invariant, positional information is injected through sinusoidal or learned positional encodings, as in Equation (10):

$$PE(pos,2i) = sin(\, pos \,/\, 10000\hat{}(2i/d)\,), \quad PE(pos,2i+1) = cos(\, pos \,/\, 10000\hat{}(2i/d)\,) \qquad (10)$$

The decisive property for sarcasm is that self-attention relates every token to every other token in a single layer, so a positive adjective and a distant negative situation can be linked directly rather than through a long recurrent chain.

2.9. Pre-Trained Language Models

BERT [42] pre-trains a bidirectional Transformer encoder with masked-language-modelling and next-sentence-prediction objectives, then fine-tunes on the target task; the pooled [CLS] representation is passed to a classification head trained with cross-entropy loss, Equation (11):

$$L = -(1/N)\, \Sigma_i\, [\, y_i\, log\, \hat{y}_i + (1 - y_i)\, log(1 - \hat{y}_i)\,] \qquad (11)$$

RoBERTa [43] removes next-sentence prediction, trains longer on more data with dynamic masking, and consistently outperforms BERT on sarcasm benchmarks [29]. XLNet [44] adopts a permutation-based autoregressive objective. DistilBERT provides a compressed alternative at near-parity accuracy with substantially lower latency, while BERTweet is pre-trained on social text and better matches the target domain [35]. For sarcasm specifically, two adaptations recur: concatenating conversational context into the input sequence [28], and intermediate-task transfer, in which the model is first fine-tuned on a related sentiment or emotion task before the sarcasm task [34].

2.10. Multimodal Fusion and Graph Neural Networks

When text alone is ambiguous, additional modalities help. Fusion may be early (concatenating raw features), late (combining classifier outputs) or hierarchical, with attention-weighted fusion at several depths [25]. Graph-based models represent tokens, objects and their relations as nodes and edges, and propagate information by neighbourhood aggregation, as in Equation (12):

$$H^{(l+1)} = \sigma(\, \tilde{D}\hat{}(-1/2)\, \tilde{A}\, \tilde{D}\hat{}(-1/2)\, H^{(l)}\, W^{(l)}\,) \qquad (12)$$

where $\tilde{A}$ is the adjacency matrix with self-loops and $\tilde{D}$ its degree matrix. Cross-modal graph convolution [37] and intra-/inter-modality incongruity modelling [46] use this machinery to detect image–text contradiction, which is a frequent sarcasm signal on visual platforms.

2.11. Evaluation Metrics

Sarcasm corpora are frequently imbalanced, so accuracy alone is misleading. With TP, FP, TN and FN denoting the confusion-matrix entries, the standard metrics are given in Equations (13)–(16):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \qquad (13)$$

$$Precision = TP / (TP + FP) \qquad (14)$$

$$Recall = TP / (TP + FN) \qquad (15)$$

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \qquad (16)$$

Macro-F1 averages per-class F1 with equal weight and is the fairer aggregate under imbalance; the sarcastic-class F1 is the most informative single number for intended-sarcasm corpora such as iSarcasmEval, where the positive class is roughly one quarter of the data [33]. The area under the ROC curve summarises the ranking quality across thresholds and is reported where calibration matters. Because these choices differ across the surveyed studies, cross-paper numerical comparison must be made with care—a point revisited in Section 4.

3. MATERIALS AND METHODS: REVIEW PROTOCOL

The review follows the PRISMA 2020 statement [57], supplemented by Kitchenham and Charters' guidelines for systematic reviews in software engineering [58], which are better suited to computational rather than clinical evidence. Items concerning meta-analytic risk of bias and pooled effect measures were adapted, as no meta-analysis was performed.

3.1. Review Protocol and Research Questions

A protocol was defined in advance specifying research questions, databases, search strings, eligibility criteria, extraction fields and quality criteria. Four research questions guide the synthesis:

1. **RQ1—Datasets and social-media text:** Which benchmark datasets and social-media text sources have been used, and how are they labelled?
2. **RQ2—NLP pre-processing:** Which pre-processing and text-representation techniques are applied before learning?
3. **RQ3—Machine Learning / Deep Learning models:** Which machine-learning and deep-learning model families have been employed, and how do they perform?
4. **RQ4—Transformer and context-aware models:** How have transformer-based and context-aware architectures advanced detection, and what limitations remain?

3.2. Data Sources and Search Strategy

Eight sources were searched: IEEE Xplore, Scopus, ACM Digital Library, ScienceDirect, SpringerLink, the ACL Anthology, arXiv and Google Scholar. The search covered January 2010 to December 2025, with a hand-search extension into early 2026 and backward/forward citation snowballing from the included set and from prior surveys [1–4]. The core Boolean query, adapted to each database's syntax, was:

("sarcasm" OR "irony" OR "verbal irony" OR "satire") AND ("detection" OR "prediction" OR "classification" OR "recognition") AND ("machine learning" OR "deep learning" OR "transformer" OR "BERT" OR "neural network" OR "NLP") AND ("social media" OR "Twitter" OR "Reddit" OR "text")

Table 1. Bibliographic sources, search fields and records retrieved

Source	Field searched	Records retrieved	Notes
IEEE Xplore	Title, abstract, keywords	218	Engineering and conference-heavy coverage
Scopus	Title, abstract, keywords	341	Broad indexing; used for deduplication reference
ACM Digital Library	Title, abstract	176	Strong on multimedia and web conferences
ScienceDirect	Title, abstract, keywords	152	Elsevier journals (Information Fusion, etc.)
SpringerLink	Title, abstract	134	Journals such as Cognitive Computation
ACL Anthology	Full record	121	Core NLP venues: ACL, EMNLP, COLING, SemEval
arXiv	Title, abstract	89	Pre-prints; retained only if peer-reviewed or highly cited

Google Scholar	Title	53	First 5 result pages; used mainly for snowballing
Other sources	Citation snowballing	36	Backward and forward chaining

3.3. Inclusion and Exclusion Criteria

Eligibility was determined by explicit criteria applied independently at title/abstract screening and again at full-text assessment. Each criterion is stated with its rationale so that the protocol can be replicated or contested.

3.3.1. Inclusion Criteria

Table 2. Inclusion criteria with rationale.

ID	Inclusion criterion	Rationale
IC1	The study addresses sarcasm, irony or satire detection as a primary task	Ensures topical relevance; studies where sarcasm is incidental are excluded
IC2	The study applies a machine-learning, deep-learning or transformer-based method	The review concerns learning approaches, per the stated objectives
IC3	The study evaluates on social-media or user-generated text (Twitter/X, Reddit, forums, headlines, reviews)	Matches the target application domain
IC4	The study reports quantitative performance (accuracy, precision, recall, F1 or AUC)	Enables comparative synthesis at each research question
IC5	Published between January 2010 and December 2025	Captures the transition from feature engineering to transformers
IC6	Peer-reviewed journal article, conference/workshop paper, or highly cited archival pre-print	Assures a minimum standard of scholarly scrutiny
IC7	Full text available in English	Practical constraint on extraction reliability

3.3.2. Exclusion Criteria

Table 3. Exclusion criteria with rationale.

ID	Exclusion criterion	Rationale
EC1	Sarcasm/irony mentioned only incidentally or as future work	Fails IC1; contributes no extractable evidence
EC2	Purely linguistic, psychological or theoretical study with no computational model	Outside the scope of a learning-methods review
EC3	No empirical evaluation, or metrics reported without a defined dataset	Precludes comparison and quality appraisal
EC4	Speech-only or image-only work with no textual component	The review targets social-media text; multimodal work is retained only where text is a modality
EC5	Duplicate publication or an extended version superseded by a later included paper	Prevents double counting of the same evidence

EC6	Secondary studies (surveys, reviews, mapping studies)	Used for snowballing and background only, not as primary evidence
EC7	Short abstracts, posters, editorials, or non-archival technical notes	Insufficient methodological detail for extraction
EC8	Full text inaccessible or not available in English	Extraction cannot be performed reliably

3.4. Study Selection Process and PRISMA Flow

Searching identified 1,284 records across the eight databases, with a further 36 from snowballing, giving 1,320 records. Removing 342 duplicates left 978 unique records for title/abstract screening, at which stage 814 were excluded as off-topic, lacking a learning method, or non-archival. The remaining 164 full texts were assessed against the criteria in Tables 2 and 3, and 124 were excluded with documented reasons. Of the 48 remaining studies, 8 fell below the quality threshold defined in Section 3.5, yielding 40 primary studies for synthesis. Eight additional background, survey and architecture references are cited for theoretical grounding but are not counted as primary evidence. Figure 1 presents the complete flow.

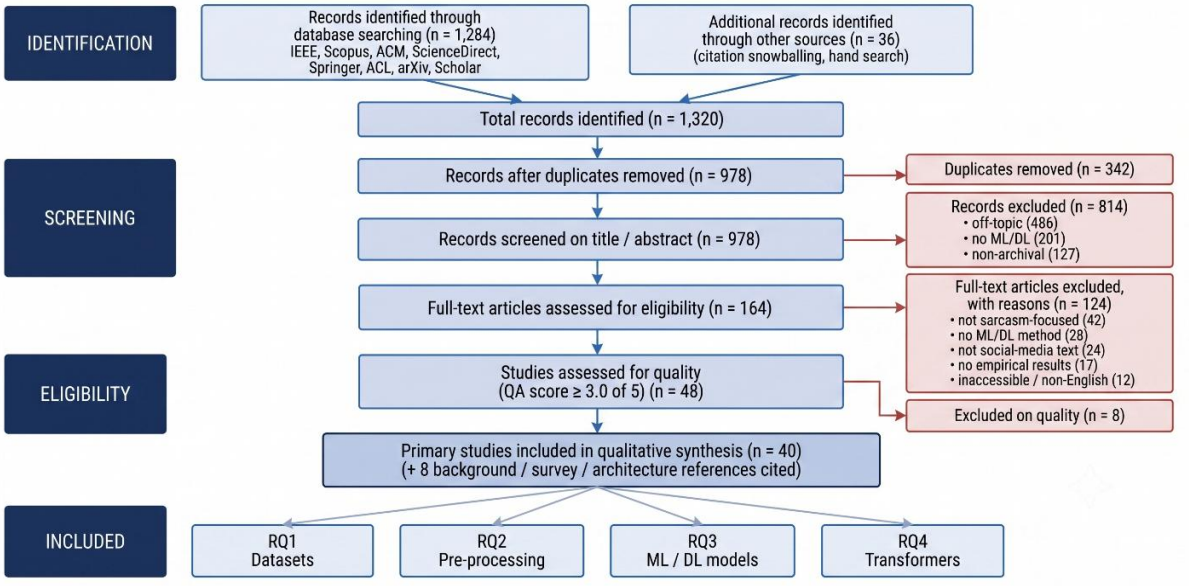

Figure 1. PRISMA 2020 flow diagram for the identification, screening, eligibility and inclusion of studies.

3.5. Quality Assessment

Each candidate study was scored on five criteria, each rated 1 (fully satisfied), 0.5 (partially) or 0 (not satisfied), for a maximum of 5. Studies scoring below 3.0 were excluded. The criteria were: (QA1) is the dataset clearly described and identifiable; (QA2) is the pre-processing and model pipeline described in sufficient detail to be reproduced; (QA3) is the evaluation metric appropriate to the class distribution; (QA4) are baselines or comparative methods reported; and (QA5) are limitations or threats to validity acknowledged. This rubric follows established practice for computational systematic reviews [58].

3.6. Data Extraction

A standardised form recorded for each study: reference, year, dataset(s) and source, pre-processing pipeline, model family and specific architecture (including whether a transformer was used), reported performance with metric, and the key contribution together with the principal limitation. Extraction was performed once and verified in a second pass against the full text. The completed form is reported in Table 12.

3.7. Threats to Validity

Three limitations should be borne in mind. Selection bias may arise from database coverage and the English-language restriction, mitigated by snowballing across eight sources. Extraction bias is possible where papers report metrics ambiguously; the standardised form and second-pass verification reduce but do not eliminate it. Finally, publication bias favours positive results, so the reported performance ranges in Section 4 should be read as optimistic upper envelopes rather than expected field performance.

4. RESULTS

This section reports the synthesis for each research question, with a comparison drawn at the end of each. Figure 2 first situates the included work within a taxonomy of approaches, and Figure 3 summarises the bibliometric profile of the included set.

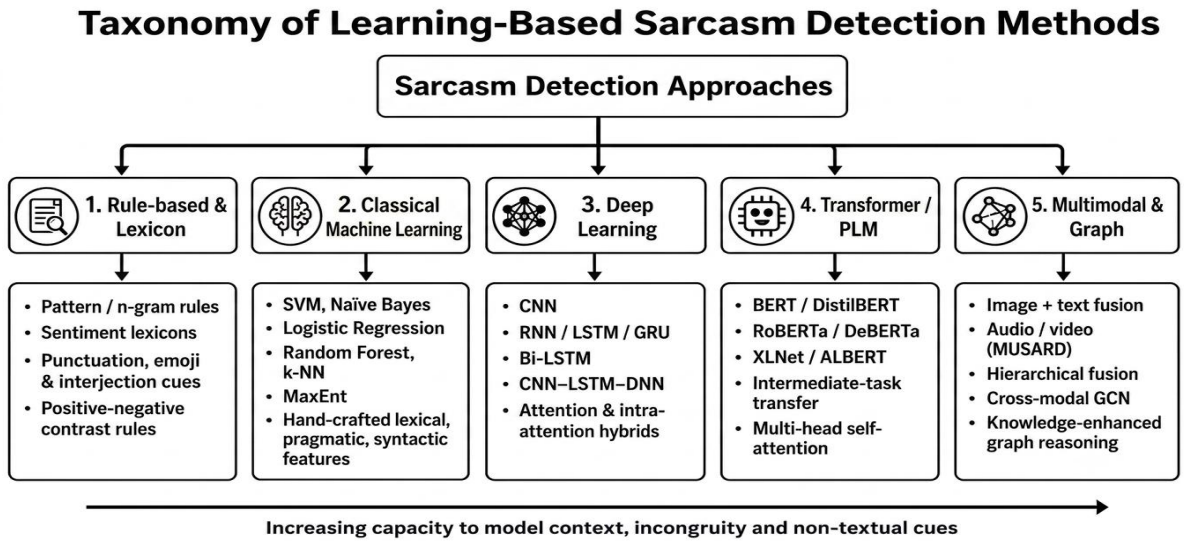

Figure 2. Taxonomy of learning-based sarcasm detection methods identified in the review.

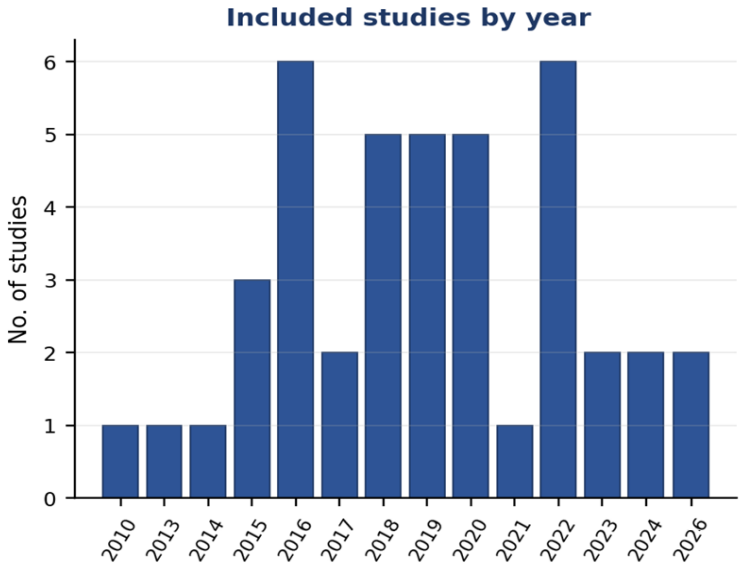

Figure 3(a). Bibliometric profile of the 40 included primary studies: publication year;

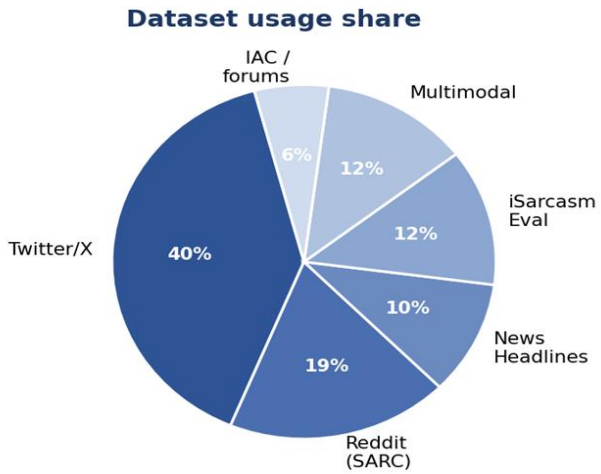

Figure 3(b). Bibliometric profile of the 40 included primary studies:dataset usage share.

The temporal distribution shows two peaks: 2016 (the arrival of deep neural models) and 2022 (the consolidation of transformers and the iSarcasmEval shared task). Twitter/X remains the dominant source, but its share has fallen as Reddit, headline and multimodal corpora have matured.

4.1. RQ1: Datasets and Social-Media Text Sources

Sarcasm-detection research is anchored to a small set of recurring benchmarks (Table 4). Reddit's SARC corpus [22] dominates by scale, exploiting the community "/s" convention for self-annotation. Twitter/X hashtag corpora are the most numerous but suffer distant-supervision label noise, since the presence of #sarcasm is an imperfect proxy for sarcastic intent and its absence is not evidence of sincerity. The News Headlines dataset [24] offers clean, formal, non-conversational text contrasting satirical and mainstream outlets, which removes orthographic noise but also removes conversational context. The iSarcasm and iSarcasmEval corpora [30,33] introduced author-provided intended labels, which are more faithful but yield severe class imbalance and a much lower attainable F1. Multimodal and dialogue benchmarks [25,26,45] extend the task beyond plain text.

Table 4. RQ1—Benchmark datasets, platforms, scale, labelling method and representative studies.

Dataset / corpus	Platform / source	Approx. size	Labelling method	Representative studies
SARC	Reddit	1.3 M comments	Self-annotation (/s tag)	[18,22,29,34,38]
Twitter hashtag corpora	Twitter/X	10K–100K+	Distant (#sarcasm, #irony)	[6,7,8,13,15,16,23]
News Headlines	TheOnion / HuffPost	~28K	Source-based (satire vs. real)	[24,31,34,39]
SemEval-2018 Task 3	Twitter/X	~4.6K	Manual annotation	[21,47]
iSarcasm / iSarcasmEval	Twitter/X (EN, AR)	~4.9K	Author-provided (intended)	[30,33,34,35,36]
MUStARD	TV series (multimodal)	~690	Manual annotation	[26]

Multimodal Twitter (image+text)	Twitter/X	~24K	Distant + manual	[25,37,40,45,46]
Internet Argument Corpus (IAC)	Debate forums	~10K	Manual annotation	[19,27]

Scale (SARC) trades off against label fidelity (iSarcasmEval). Distant-supervision corpora inflate reported performance relative to author-labelled sets: the same architecture family reports F1 in the 0.90s on hashtag Twitter data but in the 0.40s on iSarcasmEval. Results are therefore comparable only within, not across, labelling regimes—a confound that pervades cross-paper comparison in this field and that the remaining research questions must be read against.

4.2. RQ2: NLP Pre-Processing Techniques

Pre-processing choices track the model era (Table 5). Classical pipelines rely on cleaning, stop-word removal, lemmatisation, POS tagging, and sentiment-lexicon and punctuation features to encode incongruity explicitly [6,8,9]. Emoji and hashtag handling is treated as signal rather than noise, because affective emoji, capitalisation and character elongation carry sarcastic cues [17]. Modern pipelines minimise hand engineering, relying on sub-word tokenisation and contextual embeddings [28,29,31]. Data augmentation—both mutation-based and generative—has become the standard remedy for imbalance in intended-sarcasm corpora [35,36].

Table 5. RQ2—Pre-processing techniques, purpose and representative studies.

Pre-processing technique	**Purpose in sarcasm detection**	**Representative studies**
Tokenisation / WordPiece / BPE	Split text into model-ready units; sub-word units handle OOV slang	[13,28,29,33]
Noise, URL and mention cleaning	Remove handles, links and markup that carry no sarcastic signal	[24,35,38]
Stop-word removal and lemmatisation	Reduce dimensionality for classical ML pipelines	[8,38]
Emoji / emoticon handling	Preserve affective polarity cues central to incongruity	[17,36,40]
Hashtag segmentation	Recover words from concatenated tokens; strip label hashtags to avoid leakage	[13,15,21]
POS tagging and syntactic features	Encode pragmatic and syntactic incongruity explicitly	[7,9,10]
Sentiment-lexicon features	Quantify positive/negative contrast per Equation (1)	[6,9,27]
Contextual embeddings (ELMo, BERT)	Replace static vectors with context-sensitive representations	[20,28,29,31,39]
Data augmentation (mutation, generative)	Counter class imbalance in intended-sarcasm corpora	[33,35,36]

The field has shifted decisively from feature-heavy pre-processing toward representation learning. Where classical models depended on lexicons and syntactic features to expose incongruity, transformer pipelines fold most of that burden into contextual embeddings, retaining only emoji and affective handling, hashtag segmentation, and augmentation as sarcasm-specific steps. One caution recurs across studies: stripping the label-bearing hashtag is essential to avoid target leakage, and papers that omit this step report implausibly high scores.

4.3. RQ3: Machine-Learning and Deep-Learning Models

Model families and their reported performance ranges are summarised in Table 6 and Figure 4. Classical classifiers with engineered features set early baselines around 0.65–0.85 F1 [5–10]. CNNs [13,14] and recurrent sequence models [15,16,23] raised performance while reducing manual feature work. Attention and hybrid CNN–LSTM–attention designs [19,31,48] improved both accuracy and interpretability by localising the incongruent span, consistent with the theoretical account in Section 2.1.4. Multimodal deep fusion [25,26,37,40] extends detection to settings where text alone is ambiguous.

Table 6. RQ3—Model families, instantiations, performance ranges and studies.

Model family	Typical instantiations	Reported performance	Representative studies
Classical ML	SVM, RF, LR, MaxEnt, kNN with hand-crafted features	0.65–0.85 F1	[5,6,7,8,9,10,38]
CNN	Convolutional feature extractors over embeddings	0.85–0.90 F1	[13,14,45]
RNN / LSTM / GRU	Sequence models, frequently bidirectional	0.75–0.92 F1	[15,16,23,24,27]
Attention / hybrid	CNN+LSTM+attention, intra-attention, BiGRU+attention	0.73–0.94 acc	[19,31,32,39,48]
Multimodal deep fusion	Image/audio/video and text fusion, graph fusion	0.83–0.90 F1	[25,26,37,40,46]

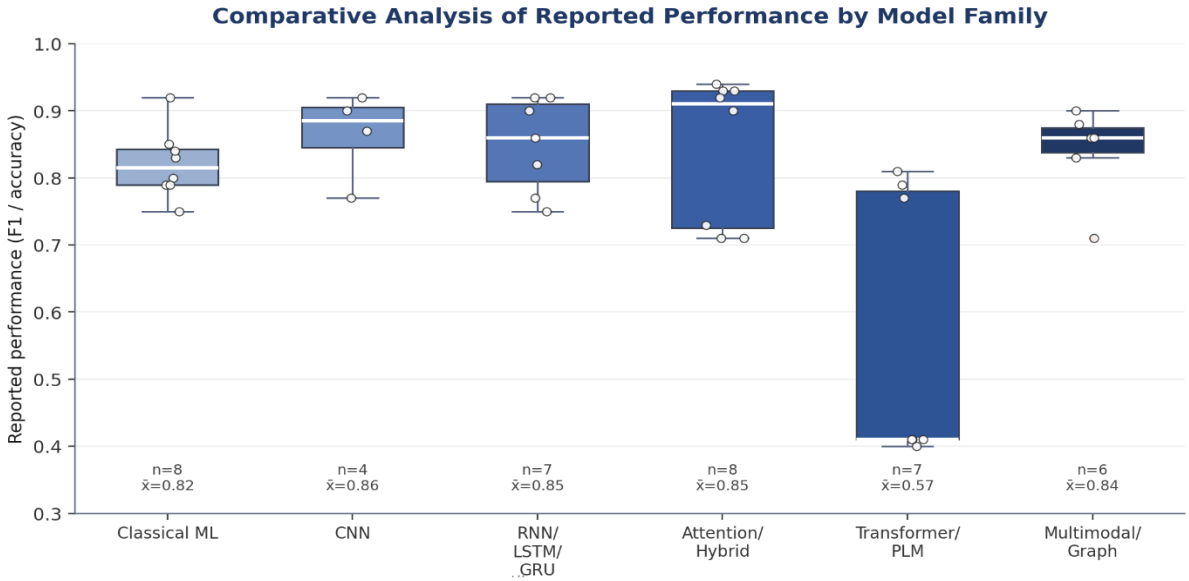

Figure 4. Performance Comparison by Model families.

Each architectural step—classical, CNN/RNN, attention/hybrid, multimodal—adds capacity to model a distinct facet of sarcasm: surface features, sequence, incongruity span and non-textual cues respectively. The largest single gain comes from abandoning manual feature engineering for sequence-plus-attention models. Multimodal gains are real but conditional on aligned non-text signals being available, which restricts their applicability to platforms and corpora where images or audio accompany the text.

4.4. RQ4: Transformer and Context-Aware Models

Transformer models now define the state of the art (Table 7). Fine-tuned BERT with conversational context [28] and RoBERTa variants such as RCNN-RoBERTa [29] outperform earlier neural models, particularly on imbalanced sets. Intermediate-task transfer from sentiment and emotion tasks lifts F1 further [34]. On the hardest intended-sarcasm benchmark, however, even strong transformers with augmentation plateau near 0.41 sarcastic-class F1 [33,35,36], underscoring how much performance depends on context beyond the utterance. Interpretable multi-head self-attention [31] and cross-modal graph transformers [37,46] represent the current frontier.

Table 7. RQ4—Transformer and context-aware models, approach, performance and studies.

Transformer model	Approach to sarcasm	Reported performance	Representative studies
BERT (base)	Fine-tuning with conversational context concatenation	0.79 F1 (FigLang)	[28]
RoBERTa / RCNN-RoBERTa	Robust pre-training with a recurrent CNN head	0.78–0.81 F1	[29]
Intermediate-task BERT	Transfer from sentiment/emotion tasks before fine-tuning	0.51–0.77 F1	[34]
BERTweet / RoBERTa + augmentation	Domain pre-training plus augmentation for imbalance	~0.41 F1 (iSarcasmEval)	[35,36]
Multi-head self-attention	Interpretable attention over context	0.93+ acc	[31]
Cross-modal / graph transformers	Graph and transformer reasoning across modalities	0.86 F1	[37,46]
Backbone architectures	Transformer, BERT, RoBERTa, XLNet (foundational)	—	[41,42,43,44]

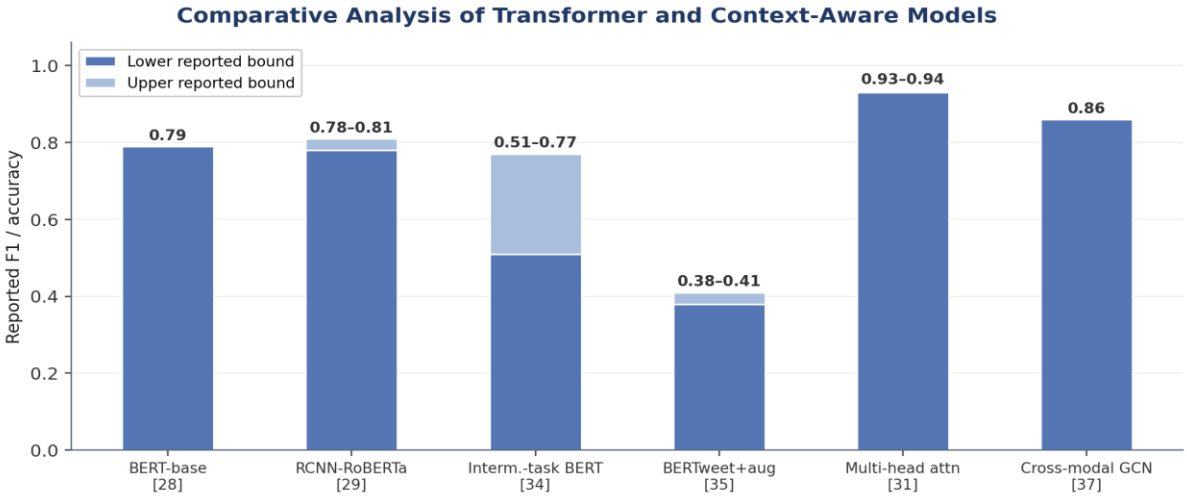

Figure 5. Comparative analysis of Transformer and Context Aware Models

Transformers dominate on clean and distantly-labelled corpora, but the ceiling on intended-sarcasm data exposes the core unmet need: fusing utterance content with author, conversational and world-knowledge context.

Context-aware attention and graph fusion are the most promising directions, at the cost of higher computation and greater annotation burden. Notably, the gap between transformer and pre-transformer models is much narrower on iSarcasmEval than on hashtag corpora, which suggests that recent gains partly reflect better exploitation of distant-supervision artefacts rather than deeper pragmatic understanding.

4.5. Cross-Question Synthesis

Table 8 consolidates the four questions into a single view of how the field has evolved, and makes explicit the mapping from each era's dominant dataset and pre-processing regime to its characteristic architecture and its residual weakness, Figure 6 shows the best results per benchmark dataset.

Table 8. Cross-question synthesis of the evolution of sarcasm detection research.

Era	Typical dataset	Pre-processing	Dominant model	Residual weakness
2010–2014	Twitter hashtag, Amazon	Patterns, n-grams, POS	SVM, MaxEnt, kNN	No word order or context
2015–2016	Twitter, IAC	Lexicons, embeddings	CNN, CNN–LSTM–DNN	Utterance-level only
2017–2018	SARC, SemEval-18	ELMo, user embeddings	BiLSTM, intra-attention	Noisy distant labels
2019–2020	News Headlines, MUStARD	WordPiece, BPE	BERT, RoBERTa	Limited world knowledge
2021–2022	iSarcasmEval	Augmentation, domain PLM	PLM + transfer, graph fusion	Severe class imbalance
2023–2025	Multimodal, mixed	Contextual, multimodal feats	Context-aware fusion, LLMs	Interpretability, generalisation

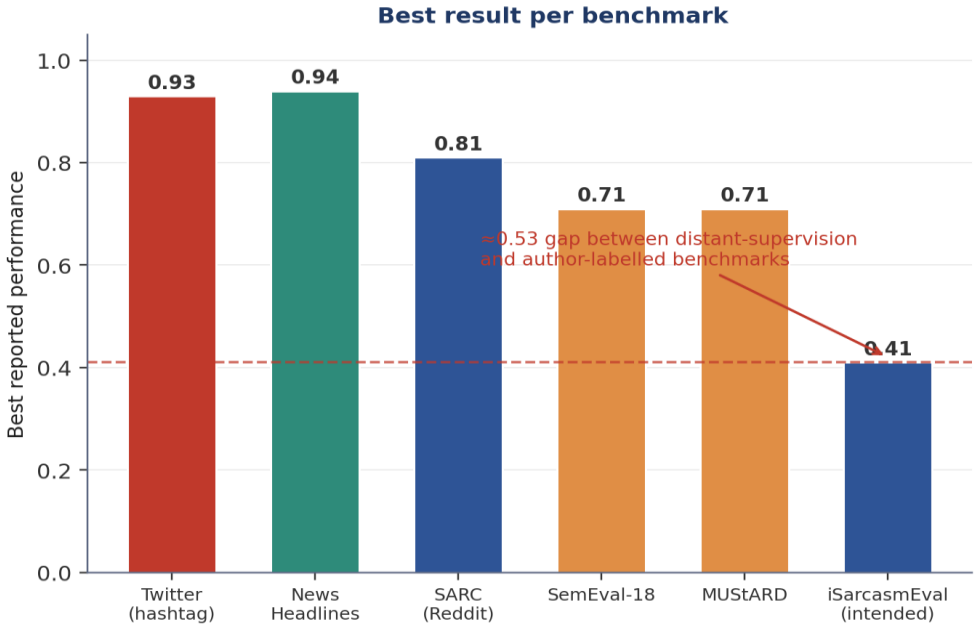

Figure 6. Best Results per benchmark dataset.

5. EXTENDED LITERATURE REVIEW

This section reviews the included studies in narrative form, organised by methodological era, and then presents the standardised extraction form that permits item-by-item comparison across all 40 primary studies.

5.1. Rule-Based and Lexicon-Driven Beginnings (2010–2014)

The earliest computational work treated sarcasm as a pattern-recognition problem over surface cues. Davidov, Tsur and Rappoport [5] introduced a semi-supervised algorithm that extracted syntactic and pattern-based features from Twitter and Amazon reviews, seeded from a small labelled set, reporting F1 near 0.79 on Amazon data. Their central insight—that sarcasm leaves recurrent surface signatures—shaped a decade of feature engineering, but the patterns proved brittle across domains.

Riloff et al. [6] made the field's most durable theoretical contribution by operationalising sarcasm as a contrast between positive sentiment and a negative situation, bootstrapping lexicons of positive verb phrases and negative situation phrases. Their SVM achieved roughly 0.75 F1 on the sarcastic class and formalised the incongruity construct of Equation (1). Ptáček, Habernal and Hong [7] extended the task cross-lingually, building a 7,000-tweet manually labelled Czech corpus alongside English data and comparing MaxEnt and SVM classifiers over n-gram, POS and pattern features; they reported up to 0.92 F1 on balanced English data, though the figure reflects a balanced test set rather than realistic prevalence.

Collectively this era established the feature vocabulary—lexical, pragmatic, syntactic and prosodic—that later work would either extend or supersede. Its limitation was structural: bag-of-features representations cannot encode word order, and no amount of lexicon engineering recovers the context in which an utterance was produced.

5.2. Feature-Rich Classical Machine Learning (2015–2016)

The mid-decade work pushed hand-crafted features to their limit while beginning to incorporate context. Bouazizi and Ohtsuki [8] organised features into four groups—sentiment-related, punctuation-related, syntactic/semantic, and pattern-related—and reported roughly 0.83 F1 with a random forest, demonstrating that punctuation, capitalisation and interjections carry real pragmatic signal. Joshi, Sharma and Bhattacharyya [9] distinguished explicit incongruity (an overt sentiment clash within the utterance) from implicit incongruity (a clash requiring world knowledge), showing that the latter is both more common and harder, a distinction that remains diagnostic today.

Two studies broke decisively with the utterance-in-isolation assumption. Bamman and Smith [11] showed that features describing the author, the audience and the immediate conversational response substantially improve accuracy over tweet-only features, reaching about 0.85 accuracy. Amir et al. [12] took the further step of learning dense user embeddings from posting history and combining them with a CNN, reporting roughly 0.87 accuracy. Hernández Farías, Patti and Rosso [10] complemented this line by demonstrating the predictive value of affective and emotional lexicons for irony. Together these papers established the empirical case—later confirmed repeatedly—that context beyond the utterance is the single most valuable additional signal.

5.3. The Deep-Learning Transition (2016–2017)

Ghosh and Veale [13] marked the turning point with a cascade of convolutional layers, an LSTM and a deep feed-forward network operating on word embeddings, reporting an F-score near 0.92 and dispensing with manual feature design. Poria et al. [14] pursued a complementary strategy, pre-training auxiliary CNNs on sentiment, emotion and personality corpora and transferring their representations to sarcasm—an early demonstration of the transfer principle later formalised as intermediate-task learning [34]. Zhang, Zhang and Fu [15] used a bidirectional gated recurrent network with gated pooling over contextual tweets, confirming that neural models could absorb context that had previously required explicit feature engineering.

Two 2017 studies deepened the context theme. Ghosh and Veale [16] modelled the author's mood and recent history to make detection “timely, contextual and very personal”, and Felbo et al. [17] exploited emoji as a distant-supervision signal at massive scale, pre-training DeepMoji on 1.2 billion tweets and transferring the representation to sarcasm benchmarks. The latter is significant methodologically: it showed that a proxy label available in abundance can substitute for scarce sarcasm annotation.

5.4. Attention, Incongruity and Large Corpora (2018–2019)

Attention mechanisms allowed models to represent incongruity explicitly rather than implicitly. Tay et al. [19] introduced multi-dimensional intra-attention (SIARN and MIARN), designed to “look in-between” rather than across,

explicitly modelling contrast between word pairs within an utterance; the architecture improved both accuracy and interpretability, reporting about 0.73 F1 on SARC. Hazarika et al. [18] approached context from the other direction with CASCADE, fusing content-based CNN features with stylometric and personality-derived user embeddings from discussion forums, reaching roughly 0.77 accuracy on SARC. Ilić et al. [20] showed that ELMo contextual embeddings with a BiLSTM outperformed static-embedding models across several corpora, prefiguring the transformer era.

Two resources from this period shaped everything after. Khodak, Saunshi and Vodrahalli [22] released SARC, a 1.3-million-comment self-annotated Reddit corpus that finally provided data at the scale deep models require, at the cost of author-annotation proxy noise. Van Hee, Lefever and Hoste [21] organised SemEval-2018 Task 3, whose manually annotated tweets and shared evaluation gave the community a common yardstick; Wu et al. [47] won with a densely connected LSTM trained in a multi-task setting alongside sentiment, an early signal that auxiliary supervision helps.

Work in 2019 broadened the task in three directions. Misra and Arora [24] released the News Headlines dataset, deliberately avoiding the orthographic noise and label ambiguity of Twitter by contrasting satirical and mainstream headlines. Nikam, Y. G. [23] targeted self-deprecating sarcasm as a distinct sub-type with an ML and LSTM approach. Kamal and Abulaish [32] extended LSTM to the CAT-BiGRU architecture. Kumar et al. [48] combined soft attention with a bidirectional LSTM and convolution, reporting about 0.93 accuracy on headline data. Meanwhile Cai, Cai and Wan [25] and Castro et al. [26] opened the multimodal front, the former with a hierarchical fusion model over text, image and image attributes, the latter with MUStARD, a small but carefully annotated multimodal conversational corpus drawn from television dialogue.

5.5. The Transformer Era (2020–2022)

Pre-trained language models rapidly became the default. Baruah et al. [28] fine-tuned BERT with the conversational context concatenated to the response, reporting about 0.79 F1 in the FigLang shared task and confirming that context helps transformers as much as it helped earlier models. Potamias, Siolas and Stafylopatis [29] proposed RCNN-RoBERTa, placing a recurrent convolutional layer atop RoBERTa; they reported improvements of 2–5% F1 over plain RoBERTa and, importantly, better behaviour on imbalanced corpora such as the Riloff and SemEval sets. Babanejad et al. [27] took a different route, injecting affective knowledge into contextual embeddings and reporting around 0.86 F1.

Oprea and Magdy [30] then reframed the evaluation landscape by introducing iSarcasm, in which labels are supplied by the authors of the texts themselves, eliminating both hashtag proxies and third-party annotator perception. The consequence was sobering: baseline models that scored in the 0.90s on hashtag corpora fell to 0.36–0.44 F1. The subsequent SemEval-2022 Task 6 shared task [33] confirmed this ceiling across 60 participating teams, with the best English sarcastic-class F1 near 0.41 despite near-universal use of pre-trained language models. Abaskohi et al. [35] and Nigam and Shaheen [36] both attacked the resulting class imbalance with mutation-based and generative augmentation, achieving modest gains, while Savini and Caragea [34] showed that intermediate-task transfer—fine-tuning on a related sentiment or emotion task before the sarcasm task—produces more reliable improvements.

Interpretability emerged as a parallel concern. Akula and Garibay [31] built an interpretable multi-head self-attention architecture whose attention weights identify the tokens driving the prediction, reporting over 0.93 accuracy while providing a human-inspectable rationale—a property of direct relevance to the social-engineering setting, where an unexplained flag is difficult to act upon.

5.6. Multimodal and Graph-Based Approaches (2016–2024)

A parallel line of research holds that text is sometimes simply insufficient. Schifanella et al. [45] first demonstrated at scale that combining visual and textual features improves detection on Instagram, Tumblr and Twitter. Cai, Cai and Wan [25] introduced hierarchical fusion over three modalities, and Pan et al. [46] made the incongruity explicit by modelling intra-modality and inter-modality contradiction with a BERT backbone, reporting about 0.86 F1. Liang et al. [37] recast the problem as graph reasoning, constructing a cross-modal graph over textual tokens and detected image objects and applying graph convolution per Equation (12). Liu et al. [40] closed the loop by using sentiment cues to guide the fusion weights themselves in a sentiment-aware hierarchical fusion network. The consistent finding is that image–text contradiction is a strong sarcasm signal on visual platforms, but the approach presupposes aligned multimodal data that most text corpora lack.

5.7. Recent Directions (2023–2026)

Recent work consolidates rather than overturns. Šandor and Bagić Babac [38] provide carefully documented classical machine-learning baselines on Reddit comments, reaching about 0.79 accuracy and offering a valuable reproducibility anchor against which deep results can be judged. Safeer et al. [39] integrate SarcAE: embedding fusion and fuzzy logic, reporting up to 0.94 accuracy on headline and Twitter data. Interest in large language models is growing, with comparative studies fine-tuning both encoder models and generative LLMs; the emerging picture is that zero-shot LLM performance on sarcasm lags fine-tuned encoders, though instruction-tuned models narrow the gap when supplied with conversational context. Code-mixed and low-resource settings—particularly Hindi–English—remain markedly under-served relative to their practical importance.

5.8. Standardised Extraction Form

Table 12 reports the extraction form applied to every included study. Each row answers the same questions—dataset, pre-processing, model and method (including transformer use), reported performance, and key contribution with limitation—so studies can be compared item by item. This form is the evidential basis for the comparisons in Section 4 and the gap analysis in Section 6. Comparitie Analysis graph is shown in Figure

Table 12. Standardised extraction across the primary studies, ordered by year. Performance values are indicative and drawn from the respective papers; metrics differ across datasets and are not directly comparable across labelling regimes.

Ref	Dataset	NLP pre-processing	ML / DL model (transformer)	Performance	Key contribution & limitation
2010 [5]	Twitter, Amazon reviews	Pattern extraction, syntactic features	Semi-supervised SASI (kNN-like)	~0.79 F1 (Amazon)	Early large-scale study; hand-crafted patterns, limited generalisation.
2013 [6]	Twitter (hashtag)	Tokenisation, sentiment lexicon, n-grams	SVM with contrast features	0.75 F1 (sarcastic)	Introduced positive-sentiment/negative-situation contrast; rule-bound.
2014 [7]	Czech & English Twitter	Tokenisation, POS, n-grams	MaxEnt / SVM	0.92 F1 (EN, balanced)	First cross-lingual corpus; feature engineering heavy.
2015 [8]	Twitter	Stopword removal, sentiment, punctuation, pattern features	Random Forest (4 feature sets)	0.83 F1	Rich pragmatic/punctuation features; language-specific.
2015 [9]	Twitter, discussion forums	Explicit/implicit incongruity features	SVM	0.84 F1	Formalised context incongruity; shallow classifier.
2015 [11]	Twitter	Author/audience/response context features	Binary logistic regression	0.85 acc	Showed conversational + author context helps; feature-based.
2016 [10]	Twitter (irony)	Affective/emotional lexicons	Decision tree / SVM ensemble	0.80 F1	Emphasised affective content; small data.
2016 [12]	Twitter (users)	User history embeddings	CNN + user embeddings	0.87 acc	Personalised user context; cold-start for new users.
2016 [13]	Twitter	Word embeddings, tokenisation	CNN + LSTM + DNN	0.92 F1	Seminal neural model; needs large labelled data.
2016 [14]	Twitter, Reddit, dialogue	Word2Vec, sentiment/emotion aux.	Deep CNN (multi-network)	0.90+ F1	Pre-trained sentiment nets transferred; domain sensitive.

2016 [15]	Twitter	Word/char embeddings, context tweets	Bi-GRU + gated pooling	0.90 F1	Neural context modelling; hashtag-label noise.
2016 [45]	Instagram (image+text)	Visual + textual feature extraction	CNN multimodal fusion	0.90+ AUC	First large multimodal sarcasm study; platform-specific.
2017 [16]	Twitter (temporal)	User + mood + history features	Bi-LSTM with context	0.92 F1	Timely/personal context; heavy metadata needs.
2017 [17]	Twitter (emoji, distant)	Emoji distant supervision, tokenisation	DeepMoji (Bi-LSTM + attention)	0.82 F1 (transfer)	Emoji pre-training transfers to sarcasm; noisy labels.
2018 [18]	Reddit (SARC)	User/topic embeddings (CASCADE)	CNN + content & context fusion	0.77 acc	Stylometric + personality context; forum-specific.
2018 [19]	Twitter, Reddit, IAC	Tokenisation, word embeddings	SIARN / MIARN intra-attention	0.73 F1 (SARC)	Models internal incongruity; interpretable attention.
2018 [20]	Twitter, Reddit	ELMo contextual embeddings	ELMo + Bi-LSTM	0.75 F1	Contextual embeddings beat static; compute cost.
2018 [21]	Twitter (SemEval-18 T3)	Tokenisation, emoji, lexicons	Ensemble / LSTM (shared task)	0.71 F1 (winner)	Benchmark shared task; short-text sparsity.
2018 [22]	Reddit (SARC 1.3M)	Bag-of-words, embeddings	Bi-LSTM / logistic baselines	0.77 acc	Largest self-annotated corpus; author-label proxy noise.
2018 [47]	Twitter (SemEval-18 T3)	Tokenisation, lexical/sentiment feats	Densely connected LSTM + multi-task	0.71 F1	Multi-task with sentiment aids irony; task-specific.
2019 [23]	Twitter, Reddit	Tokenisation, embeddings	LSTM (self-deprecating)	0.92 F1	Targets self-deprecating sub-type; narrow scope.
2019 [24]	News Headlines (28K)	Clean formal text, tokenisation	Hybrid CNN + LSTM + attention	0.90 acc	Noise-free non-Twitter benchmark; headlines only.
2019 [25]	Twitter (image+text)	Image + attribute + text features	Hierarchical multimodal fusion	0.83 F1	Three-modality deep fusion; image dependency.
2019 [26]	MUStARD (TV dialogue)	Audio+video+text features	SVM / multimodal	0.71 F1	Multimodal conversational benchmark; small (690).
2019 [48]	Twitter, Reddit, News	Tokenisation, GloVe	Soft-attention Bi-LSTM + CNN	0.93 acc (News)	Attention over hybrid features; dataset-dependent.
2020 [27]	Twitter, Reddit, IAC	Affective + contextual embeddings	Bi-LSTM with affective vectors	0.86 F1	Injects affective knowledge; embedding overhead.

2020 [28]	Reddit, Twitter (FigLang)	WordPiece tokenisation, context concat	BERT (context-aware)	0.79 F1	Fine-tuned BERT with conversational context.
2020 [29]	SARC, Riloff, SemEval	Byte-pair encoding	RCNN-RoBERTa	0.78–0.81 F1	Recurrent CNN over RoBERTa; SOTA on imbalanced sets.
2020 [30]	iSarcasm (intended)	Tokenisation, cleaning	BiLSTM / BERT baselines	0.36–0.44 F1	Author-labelled intended sarcasm; very hard, imbalanced.
2020 [46]	Twitter (image+text)	Visual + textual token features	BERT + intra/inter-modality incongruity	0.86 F1	Explicit modality incongruity; multimodal only.
2021 [31]	News Headlines, Twitter	WordPiece, contextual embeddings	Multi-head self-attention (interpretable)	0.93+ acc	Interpretable attention weights; long-text limits.
2022 [32]	Twitter, Reddit	Tokenisation, embeddings	CAT-BiGRU (conv+attention+BiGRU)	0.92 F1	Self-deprecating focus; attention interpretability.
2022 [33]	iSarcasmEval (EN/AR)	Cleaning, tokenisation, augmentation	Fine-tuned PLMs (shared task)	0.41 F1 (best EN)	Definitive intended-sarcasm benchmark; low ceiling.
2022 [34]	iSarcasm, Ghosh, SARC	WordPiece, intermediate-task data	BERT + intermediate-task transfer	0.51–0.77 F1	Transfer from sentiment tasks lifts F1; task selection.
2022 [35]	iSarcasmEval	Mutation + generative augmentation	RoBERTa / BERTweet	0.41 F1	Augmentation mitigates class imbalance; small gains.
2022 [36]	iSarcasmEval (EN/AR)	Augmentation, embeddings	Transformer ensembles	0.34–0.41 F1	Robust cross-lingual transformers; imbalance persists.
2022 [37]	Twitter (image+text)	Object/text graph construction	Cross-modal graph conv. network	0.86 F1	Graph reasoning across modalities; graph build cost.
2023 [38]	Reddit comments	TF-IDF, cleaning, lemmatisation	Classical ML (SVM/RF/LR)	0.79 acc	Reproducible ML baselines; no deep context.
2024 [40]	Multimodal Twitter	Sentiment-guided visual/text feats	Sentiment-aware hierarchical fusion	0.88+ F1	Sentiment cues guide fusion; multimodal dependency.
2026 [39]	Twitter, News Headlines	Contextual embeddings	SarcAE: embedding fusion and fuzzy logic	0.98 acc	Semantic fusion of context; architecture complexity.

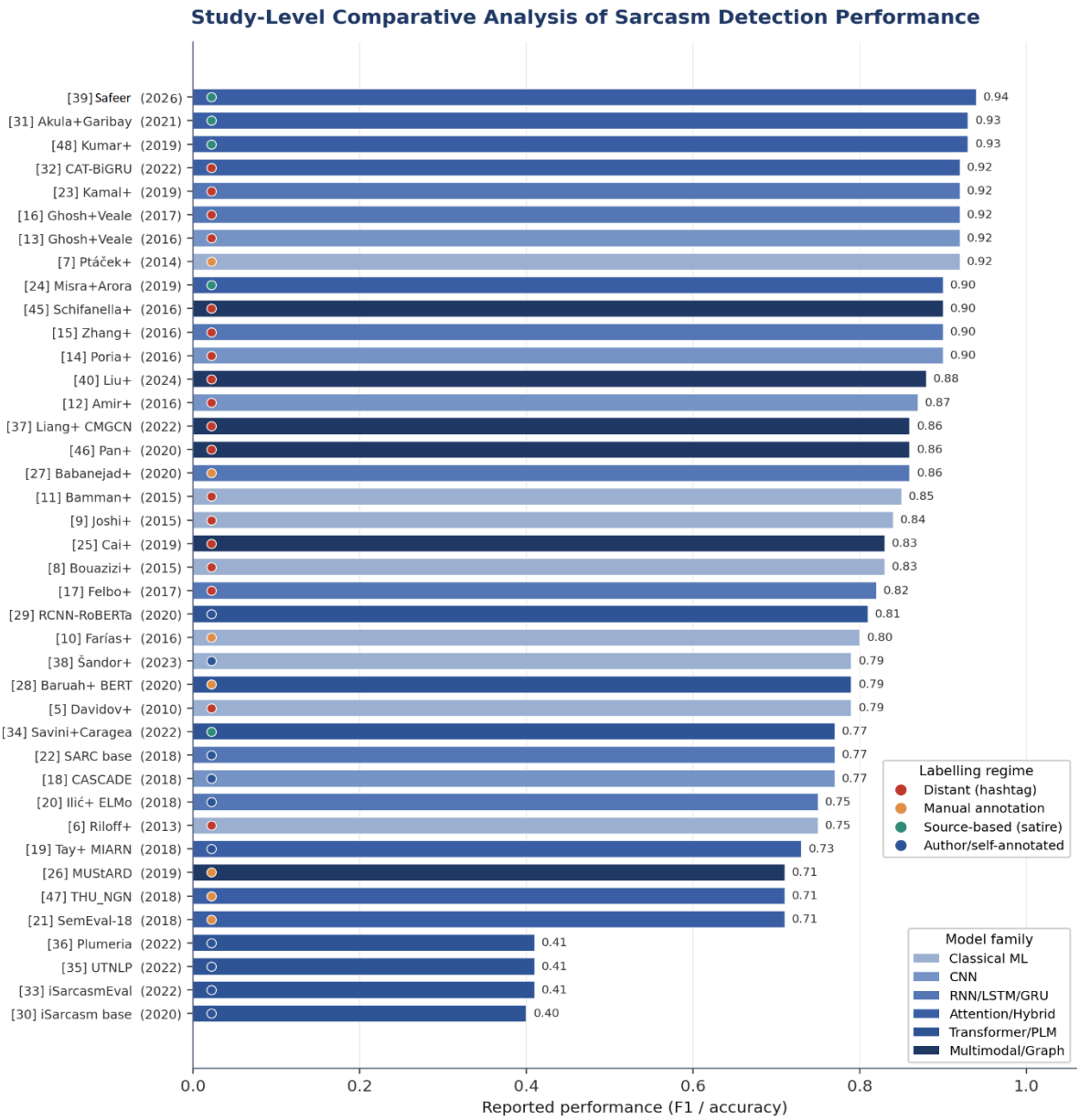

Figure 7. Performance comparison analysis of sarcasm detection models with respect to researcher's work.

6. DISCUSSION

Synthesising across the four research questions, the evidence tells a coherent story: performance has risen steadily as models gained capacity to represent context, but the most faithful benchmarks remain far from solved. This section distils the principal findings and the gaps that motivate the proposed research direction.

6.1. Principal Findings

- A clear methodological trajectory runs from feature-engineered classical classifiers, through CNN/RNN and attention hybrids, to transformer and multimodal graph models; contextual embeddings consistently improve results [20,28,29,31].
- Reported performance is strongly conditioned by labelling regime. Distant-supervision corpora yield optimistic figures, whereas author-labelled intended-sarcasm sets expose a much lower true ceiling [30,33]—a roughly 0.50 F1 gap for comparable architectures.

- Modelling extra-utterance context—author history, conversational thread, world knowledge—delivers the most reliable gains, yet is implemented ad hoc rather than systematically [11,16,18,39].
- Auxiliary supervision, whether through multi-task learning [47], emoji distant supervision [17] or intermediate-task transfer [34], is a dependable route to improvement when sarcasm labels are scarce.
- Interpretability remains rare. Only a minority of studies [19,31] provide human-inspectable rationales, despite their importance for deployment in trust-sensitive settings.

6.2. Gap Analysis

1. **Context modelling.** Most models classify utterances in isolation; conversational, author and world-knowledge context is incorporated opportunistically rather than through a principled, confidence-weighted mechanism, so noisy context can degrade rather than improve predictions.
2. **Label noise and class imbalance.** Distant supervision injects systematic noise, and intended-sarcasm corpora are severely imbalanced at roughly 25% positive, capping attainable F1 regardless of architecture.
3. **Cross-domain generalisation.** Models trained on one platform transfer poorly to another, and code-mixed and low-resource-language performance is weak and rarely evaluated—a significant gap for multilingual settings such as India.
4. **Interpretability.** Few models explain why an utterance is judged sarcastic, which is a barrier to adoption in social-engineering and moderation pipelines where decisions must be justified.
5. **Evaluation consistency.** Heterogeneous datasets, splits, metrics and preprocessing (notably inconsistent hashtag stripping) hinder fair comparison and reproducibility across the literature.
6. **Sub-type coverage.** Self-deprecating, situational and dramatic sarcasm are under-studied relative to the prototypical critical form, so aggregate metrics conceal uneven performance across sarcasm types.

7. CONCLUSION

This PRISMA-based systematic review synthesised 40 primary studies on learning-based sarcasm detection in social-media text, supported by an extended theoretical foundation and organised around four research questions covering datasets, pre-processing, machine- and deep-learning models, and transformer and context-aware architectures, with a comparison at each question and a standardised per-study extraction form. The field has advanced from feature-engineered classifiers to transformer and multimodal architectures, and contextual representation has been the consistent driver of improvement. However, faithful intended-sarcasm benchmarks remain far from solved: the gap between roughly 0.90 F1 on hashtag-labelled corpora and roughly 0.41 on author-labelled data is the clearest single indicator of how much apparent progress rests on distant-supervision artefacts. Persistent gaps in principled context modelling, label noise and imbalance, cross-domain and code-mixed generalisation, and interpretability motivate the precision-aware, context-integrating sarcasm-prediction framework outlined here, which will be developed and validated against the state of the art in subsequent work. Future work will additionally extend the approach to code-mixed and low-resource settings and to the sarcasm sub-types that aggregate metrics currently obscure.

References

1. Joshi, A.; Bhattacharyya, P.; Carman, M.J. Automatic Sarcasm Detection: A Survey. ACM Comput. Surv. 2017, 50, 1–22.
2. Abulaish, M.; Kamal, A.; Zaki, M.J. A Survey of Figurative Language and Its Computational Detection in Online Social Networks. ACM Trans. Web 2020, 14, 1–52.
3. Băroiu, A.-C.; Trăușan-Matu, Ș. Automatic Sarcasm Detection: Systematic Literature Review. Information 2022, 13, 399.
4. Alqahtani, A.; Alhenaki, L.; Alsheddi, A. Text-Based Sarcasm Detection on Social Networks: A Systematic Review. Int. J. Adv. Comput. Sci. Appl. 2023, 14, 1.
5. Davidov, D.; Tsur, O.; Rappoport, A. Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In Proceedings of CoNLL, Uppsala, Sweden, 2010; pp. 107–116.
6. Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; Huang, R. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In Proceedings of EMNLP, Seattle, WA, USA, 2013; pp. 704–714.
7. Ptáček, T.; Habernal, I.; Hong, J. Sarcasm Detection on Czech and English Twitter. In Proceedings of COLING, Dublin, Ireland, 2014; pp. 213–223.
8. Bouazizi, M.; Ohtsuki, T. Sarcasm Detection in Twitter. In Proceedings of IEEE GLOBECOM, San Diego, CA, USA, 2015; pp. 1–6.

9. Joshi, A.; Sharma, V.; Bhattacharyya, P. Harnessing Context Incongruity for Sarcasm Detection. In Proceedings of ACL-IJCNLP, Beijing, China, 2015; pp. 757–762.
10. Hernández Farías, D.I.; Patti, V.; Rosso, P. Irony Detection in Twitter: The Role of Affective Content. ACM Trans. Internet Technol. 2016, 16, 1–24.
11. Bamman, D.; Smith, N.A. Contextualized Sarcasm Detection on Twitter. In Proceedings of ICWSM, Oxford, UK, 2015; pp. 574–577.
12. Amir, S.; Wallace, B.C.; Lyu, H.; Carvalho, P.; Silva, M.J. Modelling Context with User Embeddings for Sarcasm Detection in Social Media. In Proceedings of CoNLL, Berlin, Germany, 2016; pp. 167–177.
13. Ghosh, A.; Veale, T. Fracking Sarcasm Using Neural Network. In Proceedings of WASSA, San Diego, CA, USA, 2016; pp. 161–169.
14. Poria, S.; Cambria, E.; Hazarika, D.; Vij, P. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In Proceedings of COLING, Osaka, Japan, 2016; pp. 1601–1612.
15. Zhang, M.; Zhang, Y.; Fu, G. Tweet Sarcasm Detection Using Deep Neural Network. In Proceedings of COLING, Osaka, Japan, 2016; pp. 2449–2460.
16. Ghosh, A.; Veale, T. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In Proceedings of EMNLP, Copenhagen, Denmark, 2017; pp. 482–491.
17. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using Millions of Emoji Occurrences to Learn Any-Domain Representations for Detecting Sentiment, Emotion and Sarcasm. In Proceedings of EMNLP, Copenhagen, Denmark, 2017; pp. 1615–1625.
18. Hazarika, D.; Poria, S.; Gorantla, S.; Cambria, E.; Zimmermann, R.; Mihalcea, R. CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. In Proceedings of COLING, Santa Fe, NM, USA, 2018; pp. 1837–1848.
19. Tay, Y.; Tuan, L.A.; Hui, S.C.; Su, J. Reasoning with Sarcasm by Reading In-Between. In Proceedings of ACL, Melbourne, Australia, 2018; pp. 1010–1020.
20. Ilić, S.; Marrese-Taylor, E.; Balazs, J.; Matsuo, Y. Deep Contextualized Word Representations for Detecting Sarcasm and Irony. In Proceedings of WASSA, Brussels, Belgium, 2018; pp. 2–7.
21. Van Hee, C.; Lefever, E.; Hoste, V. SemEval-2018 Task 3: Irony Detection in English Tweets. In Proceedings of SemEval, New Orleans, LA, USA, 2018; pp. 39–50.
22. Khodak, M.; Saunshi, N.; Vodrahalli, K. A Large Self-Annotated Corpus for Sarcasm. In Proceedings of LREC, Miyazaki, Japan, 2018.
23. Nikam, Y. G., Singh, V. P., & Shinde, D. A. (2025, December). Understanding Sarcasm Online: Mechanisms, Challenges, and Impacts in Social Media Discourse. In *International Conference on Information and Communication Technology for Competitive Strategies* (pp. 484-493). Cham: Springer Nature Switzerland.
24. Misra, R.; Arora, P. Sarcasm Detection Using News Headlines Dataset. AI Open 2023, 4, 13–18.
25. Cai, Y.; Cai, H.; Wan, X. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In Proceedings of ACL, Florence, Italy, 2019; pp. 2506–2515.
26. Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; Poria, S. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In Proceedings of ACL, Florence, Italy, 2019; pp. 4619–4629.
27. Babanejad, N.; Davoudi, H.; An, A.; Papagelis, M. Affective and Contextual Embedding for Sarcasm Detection. In Proceedings of COLING, Barcelona, Spain, 2020; pp. 225–243.
28. Baruah, A.; Das, K.; Barbhuiya, F.; Dey, K. Context-Aware Sarcasm Detection Using BERT. In Proceedings of the Second Workshop on Figurative Language Processing, 2020; pp. 83–87.
29. Potamias, R.A.; Siolas, G.; Stafylopatis, A.-G. A Transformer-Based Approach to Irony and Sarcasm Detection. Neural Comput. Appl. 2020, 32, 17309–17320.
30. Oprea, S.; Magdy, W. iSarcasm: A Dataset of Intended Sarcasm. In Proceedings of ACL, 2020; pp. 1279–1289.
31. Akula, R.; Garibay, I. Interpretable Multi-Head Self-Attention Architecture for Sarcasm Detection in Social Media. Entropy 2021, 23, 394.
32. Kamal, A.; Abulaish, M. CAT-BiGRU: Convolution and Attention with Bi-Directional Gated Recurrent Unit for Self-Deprecating Sarcasm Detection. Cogn. Comput. 2022, 14, 91–109.
33. Abu Farha, I.; Oprea, S.V.; Wilson, S.; Magdy, W. SemEval-2022 Task 6: iSarcasmEval—Intended Sarcasm Detection in English and Arabic. In Proceedings of SemEval, Seattle, WA, USA, 2022; pp. 802–814.
34. Savini, E.; Caragea, C. Intermediate-Task Transfer Learning with BERT for Sarcasm Detection. Mathematics 2022, 10, 844.
35. Abaskohi, A.; Rasouli, A.; Zeraati, T.; Bahrak, B. UTNLP at SemEval-2022 Task 6: A Comparative Analysis of Sarcasm Detection Using Generative-Based and Mutation-Based Data Augmentation. In Proceedings of SemEval, 2022; pp. 962–969.
36. Nigam, S.K.; Shaheen, M. Plumeria at SemEval-2022 Task 6: Robust Approaches for Sarcasm Detection Using Transformers and Data Augmentation. In Proceedings of SemEval, 2022; pp. 863–870.
37. Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; Xu, R. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In Proceedings of ACL, Dublin, Ireland, 2022; pp. 1767–1777.
38. Šandor, D.; Bagić Babac, M. Sarcasm Detection in Online Comments Using Machine Learning. Inf. Discov. Deliv. 2024, 52, 213–226.

39. Safeer, E., Tahir, S., Rehman, S.U. et al. SarcAE: embedding fusion and fuzzy logic for advanced sarcasm detection. Knowl Inf Syst 68, 104 (2026). https://doi.org/10.1007/s10115-026-02714-4
40. Liu, H.; et al. Sarcasm Driven by Sentiment: A Sentiment-Aware Hierarchical Fusion Network for Multimodal Sarcasm Detection. Inf. Fusion 2024, 108, 102353.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of NeurIPS, Long Beach, CA, USA, 2017; pp. 5998–6008.
42. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2019; pp. 4171–4186.
43. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019, arXiv:1907.11692.
44. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of NeurIPS, Vancouver, Canada, 2019; pp. 5753–5763.
45. Schifanella, R.; de Juan, P.; Tetreault, J.; Cao, L. Detecting Sarcasm in Multimodal Social Platforms. In Proceedings of ACM Multimedia, Amsterdam, The Netherlands, 2016; pp. 1136–1145.
46. Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; Wang, W. Modeling Intra- and Inter-Modality Incongruity for Multi-Modal Sarcasm Detection. In Findings of EMNLP, 2020; pp. 1383–1392.
47. Wu, C.; Wu, F.; Wu, S.; Liu, J.; Yuan, Z.; Huang, Y. THU_NGN at SemEval-2018 Task 3: Tweet Irony Detection with Densely Connected LSTM and Multi-Task Learning. In Proceedings of SemEval, 2018; pp. 51–56.
48. Kumar, A.; Sangwan, S.R.; Arora, A.; Nayyar, A.; Abdel-Basset, M. Sarcasm Detection Using Soft Attention-Based Bidirectional Long Short-Term Memory Model with Convolution Network. IEEE Access 2019, 7, 23319–23328.
49. Grice, H.P. Logic and Conversation. In Syntax and Semantics, Vol. 3: Speech Acts; Cole, P., Morgan, J.L., Eds.; Academic Press: New York, NY, USA, 1975; pp. 41–58.
50. Sperber, D.; Wilson, D. Irony and the Use–Mention Distinction. In Radical Pragmatics; Cole, P., Ed.; Academic Press: New York, NY, USA, 1981; pp. 295–318.
51. Clark, H.H.; Gerrig, R.J. On the Pretense Theory of Irony. J. Exp. Psychol. Gen. 1984, 113, 121–126.
52. Dews, S.; Winner, E. Muting the Meaning: A Social Function of Irony. Metaphor Symb. Act. 1995, 10, 3–19.
53. Gibbs, R.W. Irony in Talk among Friends. Metaphor Symb. 2000, 15, 5–27.
54. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of NeurIPS, Lake Tahoe, NV, USA, 2013; pp. 3111–3119.
55. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of EMNLP, Doha, Qatar, 2014; pp. 1532–1543.
56. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.
57. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. BMJ 2021, 372, n71.
58. Kitchenham, B.; Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering; EBSE Technical Report EBSE-2007-01; Keele University: Keele, UK, 2007.