Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.080601

REVIEW

From Lexicons to Large Language Models: A Comprehensive Survey of Sentiment Analysis Methods, Benchmarks, and Emerging Frontiers

Shuvodeep De[1,*], Agnivo Gosai[2,#], Karun Thankachan[3,#], Ramadan A. ZeinEldin[4], Abdulaziz T. Almaktoom[5], Mustafa Bayram[6] and Ali Wagdy Mohamed[7,8,*]

[1]Ingram School of Engineering, Texas State University, San Marcos, TX, USA
[2]Corning Incorporated, Painted Post, NY, USA
[3]Language Technologies Institute, School of Computer Science (SCS), Carnegie Mellon University, Pittsburgh, PA, USA
[4]Deanship of Scientific Research, King Abdulaziz University, Jeddah, Saudi Arabia
[5]Department of Operations and Supply Chain Management, Effat University, Jeddah, Saudi Arabia
[6]Department of Computer Engineering, Biruni University, Istanbul, Turkey
[7]Operations Research Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt
[8]School of Business, University of Science and Technology, Zewail City of Science and Technology, 6th of October City, Giza, Egypt

*Corresponding Authors: Shuvodeep De. Email: vvg26@txstate.edu; Ali Wagdy Mohamed. Email: aliwagdy@gmail.com

[#]These authors contributed equally to this work

Received: 12 February 2026; Accepted: 15 April 2026; Published: 27 May 2026

ABSTRACT: Sentiment analysis (SA) has evolved from a niche text-classification task into a central problem in natural language processing, spanning multiple domains, modalities, and languages. This survey provides a comprehensive review of sentiment analysis methods from their origins in lexicon-based approaches through classical machine learning, deep learning architectures, pre-trained transformers, and the current era of large language models (LLMs). We formalize the SA problem across multiple granularity levels (document, sentence, and aspect) and present a taxonomy that encompasses classification, regression, aspect-based sentiment analysis (ABSA), emotion detection, and stance detection tasks across diverse domains including movie reviews, product reviews, healthcare, finance, and social media. We review benchmark datasets spanning text-only corpora (IMDb, SST, SemEval series), multimodal benchmarks (CMU-MOSI, CMU-MOSEI, MELD), and domain-specific evaluation suites such as SentiEval. The methodological evolution is traced from VADER and SentiWordNet, through SVM and Naïve Bayes classifiers, CNN and LSTM architectures, BERT and its variants, to modern LLMs including GPT-4, Llama 3, and ModernBERT, with technical details of key architectures and their mathematical formulations. We provide dedicated analyses of chain-of-thought reasoning for implicit sentiment, multimodal fusion strategies, cross-lingual transfer methods, sarcasm and irony detection, explainability through SHAP and LIME, and the emerging challenge of AI-generated fake reviews. A comparative analysis across paradigms reveals that while LLMs achieve strong zero-shot performance, fine-tuned smaller models remain competitive on standard benchmarks, a finding with significant implications for deployment efficiency. We identify persistent open challenges including domain drift, cultural bias, and the model variability problem, and outline future research directions encompassing reasoning-augmented SA, agentic workflows, federated learning, and real-time edge deployment. With coverage of over 130 references spanning two decades of research and 29 new references from 2024 and 2025, this survey provides a unified roadmap for both newcomers and researchers at the frontier of sentiment analysis.

KEYWORDS: Sentiment analysis; opinion mining; large language models; transformers; BERT; aspect-based sentiment analysis; multimodal sentiment analysis; cross-lingual NLP; explainable AI; chain-of-thought reasoning; sarcasm detection; benchmark datasets

1 Introduction

Sentiment analysis, the computational task of identifying and extracting subjective information from text, has been a foundational problem in natural language processing (NLP) for over two decades [1,2]. What began with the pioneering 2002 work of Pang, Lee, and Vaithyanathan, who applied machine learning to movie review classification [1], has since grown into a large and widely used research and application area that touches nearly every domain where human opinion matters, including product reviews, healthcare monitoring, financial forecasting, political discourse analysis, and social media monitoring [3–5]. Over the course of its development, the field has undergone several distinct paradigm shifts, each building upon and eventually superseding the limitations of its predecessors.

The earliest systematic approaches relied on sentiment lexicons and rule-based resources such as SentiWordNet [6,7] and VADER [8], which assigned polarity scores to individual words or phrases. While interpretable and resource-efficient, these methods could not account for context-dependent sentiment, compositional semantics, or figurative language. Classical machine learning methods, including support vector machines (SVMs) [9], Naïve Bayes classifiers [10], and logistic regression, marked the first major advance by learning discriminative features directly from labeled corpora [1,11], achieving substantially higher accuracy on benchmark tasks. The subsequent deep learning revolution introduced convolutional neural networks (CNNs) [12], recurrent architectures including long short-term memory (LSTM) networks [13], and attention mechanisms [14,15], all of which could capture sequential dependencies and long-range context without manual feature engineering. A more fundamental shift arrived with transfer learning through pre-trained transformers such as BERT [16], RoBERTa [17], and XLNet [18], which changed the landscape by enabling fine-tuning on downstream sentiment tasks with minimal labeled data [19]. Most recently, large language models (LLMs) such as GPT-3 [20], GPT-4-class models, Llama 3 [21], and their instruction-tuned variants have shown strong zero-shot and few-shot sentiment analysis performance on a range of benchmarks [22–24], while simultaneously raising new questions about reliability, cost, and interpretability [25]. Fig. 1 provides a high-level overview of this methodological evolution, situating the major sentiment analysis paradigms that frame the structure of the remainder of this survey.

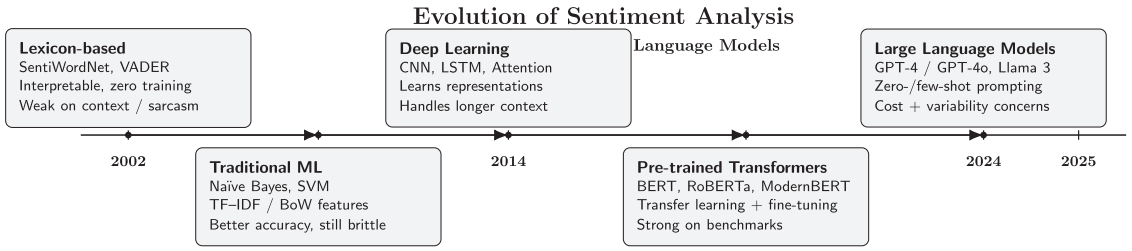

Figure 1: Evolution of sentiment analysis methodologies from early lexicon-based approaches to modern large language models. The timeline highlights major paradigm shifts, representative model families, and their characteristic strengths and limitations (This figure was generated with the assistance of OpenAI GPT 5.2).

Despite two decades of sustained progress, sentiment analysis remains far from solved. Sarcasm and irony continue to confound even state-of-the-art systems [26–28], because the intended meaning is often the polar opposite of what is literally expressed. Domain and temporal drift remain persistent sources of performance degradation, as models trained in one context, whether a product category or a time period, frequently fail to generalize to others [29,30]. The challenge of cross-lingual transfer to low-resource languages, where labeled training data is scarce or nonexistent, has attracted growing attention but remains largely unsolved [31–33]. Multimodal fusion across text, audio, and video introduces additional complexity, requiring models to integrate fundamentally different signal types into a coherent sentiment judgment

[34–36]. Furthermore, the proliferation of LLMs has given rise to an entirely new threat: the generation of sophisticated AI-generated fake reviews that can evade traditional detection systems [37,38]. Adding to these concerns, recent research has identified a *model variability problem* (MVP) in LLM-based sentiment analysis, whereby identical inputs can produce inconsistent classifications across runs due to stochastic inference and prompt sensitivity [25], a finding with troubling implications for deployment in high-stakes applications.

This survey aims to provide a comprehensive and technically grounded review of the sentiment analysis landscape, from its origins to its current frontiers. Our approach is distinguished by five core contributions that collectively provide a unifying framework absent from prior surveys: (1) a formal taxonomy (Section 2) that organizes sentiment analysis along orthogonal dimensions of granularity, task formulation, domain, and output space, serving as a scaffold that connects all subsequent methodological discussion; (2) mathematical formulations for key architectures across all five paradigms, enabling readers to understand not just *what* works but *why*; (3) coverage of 29 new references from 2024 and 2025; (4) substantive treatment of topics largely absent from existing surveys, including the model variability problem, agentic SA workflows, and AI-generated fake review detection; and (5) a structured practitioner decision framework for choosing among paradigms based on specific engineering constraints.

Several prior surveys have made valuable contributions to the systematic understanding of sentiment analysis. Wankhade et al. [4] surveyed core methods, applications, and challenges, while Jain et al. [39] conducted a systematic review of ML methods for consumer sentiment analysis in hospitality and tourism reviews. Raghunathan and Saravanakumar [40] cataloged persistent methodological challenges, and Tetteh and Thushara [41] focused specifically on evaluation tools for movie review sentiment analysis. Subsequent surveys expanded the scope to include deep learning and hybrid approaches, notably those by Birjali et al. [5], Islam et al. [42], Bordoloi and Biswas [43], and Mao et al. [44].

More recently, a new generation of surveys has emerged reflecting the rapid evolution of sentiment analysis in the post-transformer and generative AI era. Kumar et al. [45] provide a comprehensive review tracing the progression from classical machine learning techniques to transformer-based architectures, while Alahmadi et al. [46] examines generalization challenges and emerging trends across application domains. Suryawanshi [47] surveys machine learning and deep learning techniques with an emphasis on practical applications, and Bachate and Suchitra [48] focuses on sentiment and emotion recognition in social media contexts. In parallel, Krugmann and Hartmann [23] discusses the impact of generative AI on sentiment analysis workflows, and recent technical reports [49] systematically catalog datasets, tools, and evaluation challenges. Ahmad Alomari [50] presented a comprehensive PRISMA-guided survey evaluating ChatGPT across multiple NLP tasks, highlighting its adaptability, performance trends, and key limitations in real-world applications. Despite these contributions, existing surveys typically address large language models only in isolation or as an extension of prior paradigms, motivating the need for a unified survey spanning the full methodological evolution from lexicon-based methods to modern LLM-centric sentiment analysis.

However, the present survey distinguishes itself from these prior works in several critical respects. First, unlike surveys and modeling works limited to movie reviews [41,51] or single methodological families, we span the complete pipeline from lexicons to reasoning-augmented LLMs across multiple domains, including movies, products, healthcare, finance, and social media. Second, we provide mathematical formulations for key architectures, from the SVM dual formulation and LSTM gating equations to the transformer self-attention mechanism and LoRA adaptation, enabling readers to understand not just *what* works but *why*. Third, we have studied and discussed several new references from 2024 and 2025, covering LLM benchmarking [22,24,52,53], chain-of-thought reasoning for sentiment [54–57], ModernBERT [58], multimodal LLMs [36,59,60], cross-lingual methods [33,61–65], sarcasm detection with LLMs [28,66,67], explainable SA [68–70], AI-generated fake reviews [37,38], and ensemble strategies [71]. Fourth, we dedicate

substantive treatment to topics that are largely absent from existing surveys, including AI-generated fake review detection, the model variability problem, agentic SA workflows, and explainability.

The remainder of this paper is organized as follows. Section 1.1 describes our survey methodology. Section 2 formalizes the sentiment analysis problem and presents a taxonomy of tasks, granularity levels, and domains. Section 3 reviews benchmark datasets and evaluation metrics. Section 4 traces the methodological evolution from lexicons through LLMs, providing technical details at each stage. Section 5 examines domain-specific challenges including sarcasm detection, domain drift, and AI-generated fake reviews. Section 6 surveys emerging frontiers such as reasoning-augmented SA, agentic workflows, federated learning, and explainability. Section 7 identifies research gaps and future directions. Finally, Section 8 concludes the survey.

1.1 Survey Methodology

To ensure comprehensive and reproducible coverage, this survey follows a structured narrative synthesis methodology. We searched five major academic databases: Scopus, Web of Science, IEEE Xplore, ACL Anthology, and Google Scholar. Primary search queries included combinations of the terms "sentiment analysis," "opinion mining," "aspect-based sentiment analysis," "large language models AND sentiment," "multimodal sentiment," "cross-lingual sentiment," "sarcasm detection," and "explainable sentiment analysis." Searches were conducted iteratively with special focus in the time-span: October 2024 and February 2026.

Inclusion criteria required that papers be: (1) published in peer-reviewed venues or established preprint servers (arXiv) with demonstrable community impact; (2) directly relevant to sentiment analysis methodology, evaluation, or application; and (3) available in English. We excluded: purely application-specific case studies without methodological contribution, duplicate publications, and works superseded by later versions from the same authors. For rapidly evolving topics (LLM evaluation, multimodal SA, cross-lingual transfer), we prioritized publications from 2024–2025 to ensure currency.

Starting from an initial pool of over 300 candidate papers identified through database searches, we applied snowball sampling by tracing citation networks to identify seminal works and recent extensions. After applying inclusion/exclusion criteria and removing duplicates, the final survey covers 140 references covering the most recent developments. While this survey follows a narrative synthesis approach rather than a strict PRISMA systematic review protocol which is better suited to meta-analyses of empirical studies the structured search and selection process ensures breadth and reproducibility.

2 Problem Formulation and Taxonomy

To ground this historical and methodological overview in a precise analytical framework, it is necessary to formalize what constitutes a sentiment analysis task and how different problem formulations relate to one another. While the term "sentiment analysis" is often used broadly, it encompasses a diverse family of tasks that vary in granularity, output space, and application context. The taxonomy presented in Fig. 2 serves as the unifying framework for this survey: each methodological paradigm discussed in Section 4 can be understood as a different parameterization of the formal problem defined here, and each domain-specific challenge in Section 5 corresponds to a specific failure mode within one or more dimensions of this taxonomy. The next section therefore introduces a formal problem definition and a unified taxonomy that will serve as a reference point for the methodological and empirical discussions that follow.

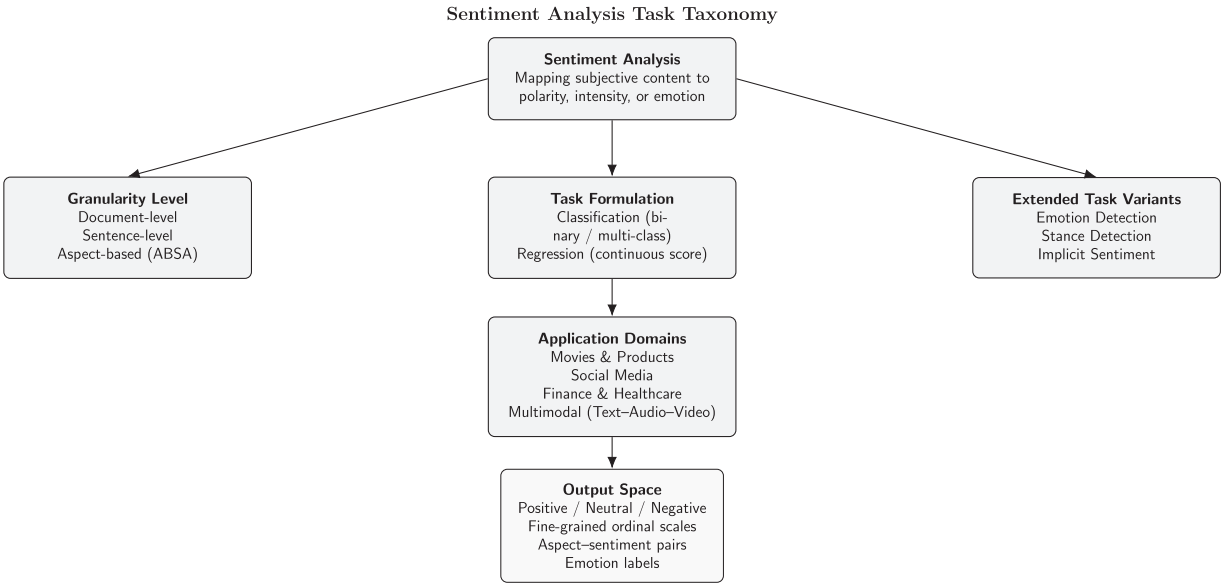

Figure 2: Taxonomy of sentiment analysis tasks. The diagram organizes sentiment analysis along orthogonal dimensions including granularity level, task formulation, extended task variants, application domains, and output spaces (This figure was generated with the assistance of OpenAI GPT 5.2).

2.1 Formal Problem Definition

Sentiment analysis most commonly classifies text into positive, negative, or neutral categories, although many benchmarks use fine-grained polarity scales such as five-class sentiment (very negative, negative, neutral, positive, very positive). Thus, we can formally define sentiment analysis can be formally defined as a mapping function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ denotes the input space (text, audio, video, or their combination) and $\mathcal{Y}$ denotes the output space of sentiment labels or scores. The nature of $\mathcal{Y}$ varies with the task formulation: in binary classification, $\mathcal{Y} = \{\text{positive}, \text{negative}\}$; in ternary classification, a neutral class is added so that $\mathcal{Y} = \{\text{positive}, \text{neutral}, \text{negative}\}$ or a multi-class polarity scale; fine-grained settings expand this further to ordinal scales such as $\mathcal{Y} = \{1, 2, 3, 4, 5\}$; regression formulations define continuous output spaces such as $\mathcal{Y} = [-1, 1]$ or $\mathcal{Y} = [0, 1]$; and aspect-based formulations condition the mapping on a specific aspect $a \in \mathcal{A}$, yielding $f : (\mathcal{X}, a) \to \mathcal{Y}$.

For a document d consisting of a sequence of tokens $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the goal is to learn parameters $\boldsymbol{\theta}$ that maximize the conditional likelihood:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log P(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) \tag{1}$$

where $\mathbf{x}_i$ denotes the input instance (e.g., a sequence of tokens representing a document or sentence), y_i is the corresponding sentiment label or score, and θ represents the model parameters and N is the number of training examples. The summation defines the log-likelihood objective, and θ^* denotes the optimal model parameters that maximize this objective. The conditional probability $P(y_i \mid \mathbf{x}_i; \theta)$ is parameterized differently depending on the model family, ranging from linear classifiers to deep neural networks. This objective corresponds to maximizing the likelihood of the observed training data under the model, i.e., encouraging the model to assign high probability to correct sentiment labels. Taking the logarithm converts the product of probabilities into a sum, improving numerical stability and simplifying optimization. This formulation also highlights a key tension that recurs throughout the survey: supervised methods optimize Eq. (1)

directly but require labeled data, while LLM-based approaches approximate the same mapping through in-context learning without explicit parameter optimization on task-specific data, trading annotation cost for computational cost and introducing the model variability problem discussed in Section 4.5.4.

2.2 Granularity Levels

Sentiment analysis is typically studied across multiple granularity levels: document-level (overall polarity), sentence-level, target-level (sentiment toward specific entities), and aspect-level (polarity toward specific entity attributes), each progressively narrowing the analytical focus [3,4,72]. At the **document level**, a single sentiment label is assigned to an entire document such as a movie review [1,73]. This formulation assumes a single dominant opinion, an assumption that breaks down for documents discussing multiple entities or aspects. **Sentence or Phrase-level** SA addresses this limitation by classifying individual sentences, recognizing that a document may contain mixed sentiments [74,75]. Wilson et al. [74] demonstrated that contextual polarity at the phrase level often diverges from prior word-level polarity, underscoring the importance of finer-grained analysis.

The most detailed formulation is **aspect-level** SA, also known as aspect-based sentiment analysis (ABSA), identifies sentiment toward specific attributes or aspects of an entity mentioned in the text (e.g., food quality, service, price) [76–80]. For example, in the sentence "The cinematography was stunning but the plot was predictable," the aspect *cinematography* carries positive sentiment while *plot* carries negative sentiment. Formally, given input $\mathbf{x}$ and aspect term a, ABSA solves:

$$\hat{y}_a = \arg\max_{y \in \mathcal{Y}} P(y \mid \mathbf{x}, a; \boldsymbol{\theta}^*) \tag{2}$$

here, $\mathcal{Y}$ denotes the set of possible sentiment labels (e.g., positive, negative, neutral). At inference time, the trained parameters $\boldsymbol{\theta}^*$ are fixed and used to predict the most likely sentiment label.

Recent work has further extended this to sub-aspect analysis [79] and cross-lingual ABSA [33], where aspect-sentiment pairs must be identified across languages with limited target-language supervision. Fig. 3 illustrates how the same text can yield substantially different sentiment interpretations depending on the chosen granularity level.

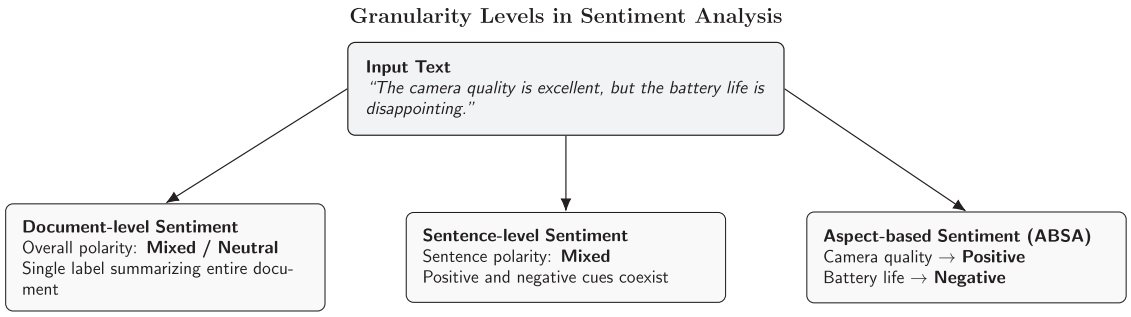

Figure 3: Illustration of sentiment analysis granularity levels. The same input text yields different sentiment interpretations depending on whether sentiment is modeled at the document, sentence, or aspect level (This figure was generated with the assistance of OpenAI GPT 5.2).

2.3 Task Variants

Beyond polarity classification, the sentiment analysis ecosystem encompasses several related tasks [4,43,44]. **Emotion detection** extends binary polarity to multi-label emotion taxonomies such as Plutchik's wheel or the GoEmotions taxonomy with 27 emotion categories [81], where the output space

becomes $\mathcal{Y} = 2^{\mathcal{E}}$ with $\mathcal{E}$ denoting the set of emotions. **Stance detection** determines the author's position (favor, against, neutral) toward a specific target, which may not be explicitly mentioned in the text [82]. **Multimodal SA** extends the input space to $\mathcal{X} = \mathcal{X}_t \times \mathcal{X}_a \times \mathcal{X}_v$, encompassing textual, acoustic, and visual modalities [34–36,83]; Yang et al. [36] provide a comprehensive survey of how LLMs are being integrated into text-centric multimodal sentiment analysis pipelines. Finally, **implicit sentiment analysis** addresses cases where sentiment is conveyed indirectly through implications, metaphors, or pragmatic inference rather than explicit opinion words [54,55], a task that has gained renewed attention through chain-of-thought prompting approaches.

2.4 Domain Taxonomy

While movie reviews served as the canonical testbed for sentiment analysis [1,41,51,84,85], the field now spans diverse domains with distinct linguistic characteristics. Product reviews on platforms such as Amazon and Yelp feature aspect-heavy text with domain-specific vocabulary [77]. Social media platforms including Twitter/X and Reddit present challenges of informal language, hashtags, emojis, sarcasm, and character limits [8,82,86]. Financial text, including earnings calls, news articles, and analyst reports, requires domain-adapted models; Shen and Zhang [52] reported that GPT-4-class models using few-shot prompting can approach the performance of fine-tuned financial sentiment models such as FinBERT. Bhatia et al. [87] later introduced FinTral, a domain-adapted financial language model based on the Mistral architecture that achieved competitive performance on financial sentiment benchmarks. Healthcare text, encompassing drug reviews, clinical notes, and patient feedback, demands high accuracy and interpretability given the stakes involved. Cross-lingual settings require transferring SA capabilities across languages, particularly to low-resource languages where labeled data is scarce [31,32,61–64]. A closely related and practically important application domain is **online hate detection and toxic content analysis**, which shares significant methodological overlap with sentiment analysis while introducing distinct challenges. Hate speech detection requires distinguishing between negative sentiment (legitimate criticism) and harmful content (targeted abuse), often in the presence of implicit toxicity, code-switching, and platform-specific linguistic norms. The TweetEval benchmark [82] includes offensive language detection as one of its unified tasks, and Ranasinghe and Zampieri [88] demonstrated that cross-lingual embeddings can transfer offensive language identification capabilities across languages, an approach directly applicable to multilingual sentiment analysis. The intersection of sentiment analysis with content moderation represents a high-stakes practical frontier where classification errors carry significant social consequences.

The diversity of task formulations, granularity levels, and domains outlined above directly shapes how sentiment analysis systems are evaluated. Models optimized for document-level polarity may perform poorly on aspect-based or implicit sentiment tasks, and benchmarks designed for one domain often fail to reflect challenges present in others. Consequently, the choice of datasets and evaluation metrics plays a decisive role in interpreting reported performance. The following section surveys the benchmark datasets and evaluation protocols that have driven progress in sentiment analysis research.

3 Benchmark Datasets and Evaluation

Benchmark datasets and evaluation protocols play a central role in shaping progress in sentiment analysis. Beyond serving as performance indicators, benchmarks implicitly define task boundaries, influence model design choices, and determine which linguistic phenomena are emphasized or overlooked. As sentiment analysis has evolved from document-level polarity classification to fine-grained, multilingual, and multimodal tasks, benchmark datasets have likewise diversified in scope, annotation strategies, and evaluation metrics. This section surveys the most widely used benchmark datasets and evaluation methodologies,

highlighting how they reflect underlying task formulations and expose both the strengths and limitations of existing sentiment analysis models. Fig. 4 situates widely used sentiment analysis benchmarks within a historical timeline, highlighting shifts in task formulation and dataset scale.

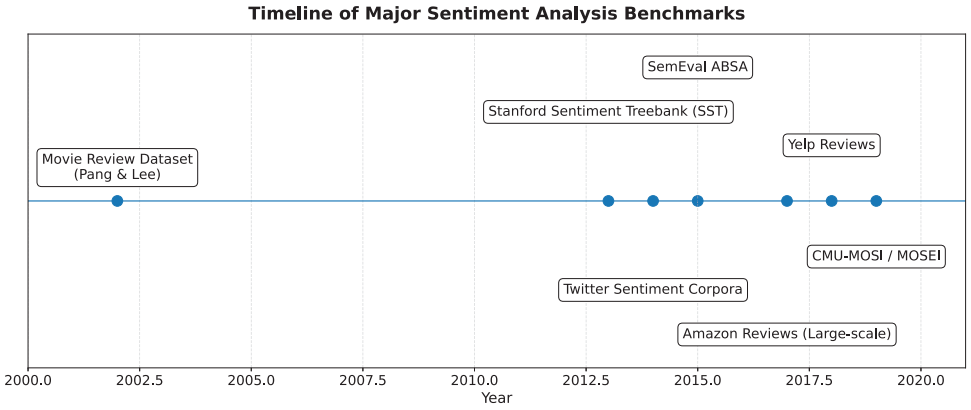

Figure 4: Timeline of major benchmark datasets in sentiment analysis, illustrating the progression from early document-level datasets to aspect-based, large-scale industrial, and multimodal sentiment benchmarks (This figure was generated with the assistance of OpenAI GPT 5.2).

3.1 Text Benchmarks

The development of standardized benchmarks has been instrumental in driving progress in sentiment analysis, and Table 1 summarizes the major text-based datasets that have shaped the field. **The Movie Review (MR)** Corpus, introduced by Pang and Lee [1,89], contains 2000 movie reviews (later expanded to 10,662) labeled as positive or negative; despite its age, it remains frequently used due to its balanced design and linguistic richness. The **IMDb Dataset** [85] provides 50,000 movie reviews with binary sentiment labels and has become the de facto standard for evaluating document-level SA, with reviews that are substantially longer than typical benchmark texts, thereby testing models' ability to handle extended context.

Table 1: Summary of major sentiment analysis benchmark datasets. These are widely used benchmarks in English language (cross-lingual research discussed in Sections 5.4 and 6.6).

Dataset	Year	Approx. Size	Classes	Domain	Key Feature
MR Corpus [1]	2002	10.6K	2	Movies	Pioneering SA dataset
IMDb [85]	2011	50K	2	Movies	Long reviews, balanced
SST-2/5 [84]	2013	11.8K	2/5	Movies	Phrase-level annotations
Amazon [29]	2007	8K	2	Products	Multi-domain transfer
SemEval-2014	2014	6.5K	3	Rest./Lap.	Aspect-based SA
GoEmotions [81]	2020	58K	28	Reddit	Fine-grained emotion
TweetEval [82]	2020	124K	Varies	Twitter	Unified tweet tasks
Fin. PhraseBank	2014	4.8K	3	Finance	Annotator agreement
SentiEval [22]	2024	26 sets	Varies	Multi	LLM evaluation suite
CMU-MOSI [90]	2016	2.2K	7	Videos	Multimodal (T+A+V)
CMU-MOSEI [83]	2018	23.4K	7	Videos	Large-scale multimodal

The **Stanford Sentiment Treebank (SST)** [84] represents a landmark contribution: Socher et al. annotated parse trees of 11,855 sentences at every node, enabling both binary (SST-2) and fine-grained five-class (SST-5) evaluation. SST-2 has become one of the most widely reported benchmarks, with BERT achieving 94.9% accuracy [16]. The **SemEval Series** of workshops has produced a rich collection of SA benchmarks, including aspect-based SA datasets (SemEval-2014 Task 4) [27], stance detection (SemEval-2016 Task 6), and figurative language sentiment [27]. More recently, Demszky et al. [81] introduced **GoEmotions**, comprising 58,009 Reddit comments annotated with 27 emotion categories plus neutral, which enables fine-grained emotion detection beyond simple polarity. Barbieri et al. [82] unified seven tweet classification tasks, including sentiment, emotion, and offensive language detection, into a single **TweetEval** benchmark.

While the historical prominence of movie review datasets (MR, IMDb, SST) reflects the field's origins, practitioners should note that the linguistic characteristics of movie reviews: lengthy, grammatically correct, opinion-focused differ substantially from the short, noisy, informal text encountered in social media, financial, and healthcare domains. This discrepancy motivates the continued development of domain-specific benchmarks discussed below.

3.2 Multimodal Benchmarks

Multimodal sentiment analysis requires datasets that align textual, acoustic, and visual signals [34,91]. **CMU-MOSI** [90] contains 2199 video segments from YouTube opinion videos annotated with sentiment intensity on a $[-3, +3]$ scale, while **CMU-MOSEI** [83] provides a substantially larger collection of 23,453 annotated segments across 1000 speakers and 250 topics, with annotations for both sentiment and six basic emotions. The **Multimodal EmotionLines Dataset (MELD)** [92] offers 13,708 utterances from the television series *Friends*, annotated with seven emotion labels and three sentiment labels, thereby enabling research on conversational emotion recognition. A recent systematic review [60] analyzing 116 multimodal SA studies from 2018–2025 identifies persistent challenges in temporal alignment between modalities and the continued dominance of text over audio-visual features in most fusion architectures.

3.3 Domain-Specific Benchmarks

Several benchmarks target specific domains or evaluation objectives. Zhang et al. introduced **SentiEval** [22], a comprehensive benchmark encompassing 13 SA tasks across 26 datasets, specifically designed to evaluate LLM capabilities. Their evaluation revealed that fine-tuned small language models (SLMs) outperform LLMs on most standard tasks, though LLMs demonstrate superior few-shot learning. The **Financial PhraseBank** contains 4845 sentences from financial news annotated by domain experts with three-class sentiment labels; Shen and Zhang [52] used this benchmark to demonstrate that GPT-4o with few-shot prompting achieves performance comparable to fine-tuned FinBERT. In the area of figurative language, Zhang et al. introduced **SarcasmBench** [28], the first comprehensive sarcasm detection benchmark for evaluating LLMs, which revealed that GPT-4 achieves 14% higher accuracy than other LLMs but that all LLMs underperform fine-tuned pre-trained language models (PLMs). Fig. 5 illustrates the diminishing performance gains observed as sentiment analysis datasets scale to increasingly large sizes.

3.4 Evaluation Metrics and Their Limitations

Standard evaluation metrics for sentiment analysis include accuracy, precision, recall, F1-score (macro and weighted), and, for regression tasks, mean absolute error (MAE) and Pearson correlation. For binary classification, accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP (true positives) and TN (true negatives) denote the number of correctly classified positive and negative instances, respectively, while FP (false positives) and FN (false negatives) represent misclassified instances. These quantities are defined with respect to a given positive class in binary classification settings. It is worth mentioning that in multi-class settings, accuracy is computed analogously as the proportion of correctly classified instances over the total number of samples.

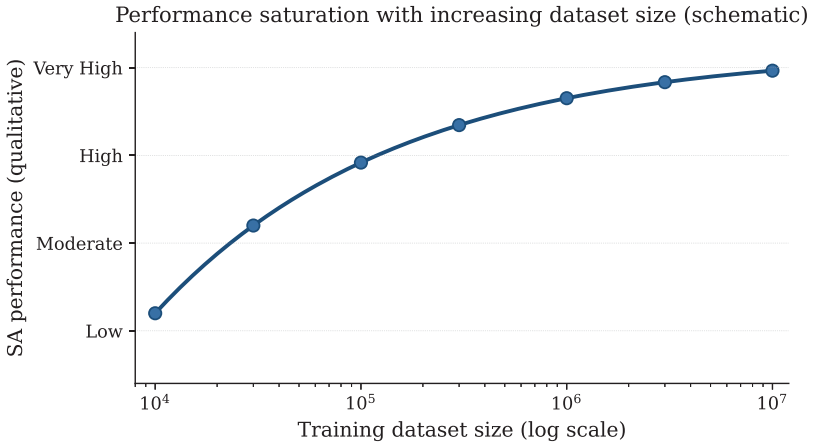

Figure 5: Illustrative relationship between training dataset size and relative sentiment analysis performance (%). The curve highlights diminishing returns as dataset scale increases, reflecting saturation trends reported across multiple benchmarks rather than results from a single experimental setup (This figure was generated with the assistance of OpenAI GPT 5.2).

Precision and recall provide complementary class-specific information:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{4}$$

where TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives for class c, respectively, computed under a one-vs-rest formulation in multi-class classification. $c \in \mathcal{Y}$ corresponds to a sentiment class such as positive, negative, or neutral.

In the SA context, the practical interpretation of these metrics varies by application: high precision for negative sentiment is critical in brand monitoring (minimizing false alarms), while high recall for negative sentiment matters more in mental health surveillance (ensuring no distress signal is missed).

To obtain a single performance measure across all classes, precision and recall are typically aggregated over classes using averaging strategies such as macro-averaging or weighted averaging. The harmonic mean of precision and recall yields the F1-score:

$$\text{F1}_{\text{macro}} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{5}$$

where P_c and R_c denote the precision and recall for class c as defined in Eq. (4). Macro-averaging assigns equal weight to each class regardless of class frequency, making it particularly suitable for imbalanced datasets where minority classes are of interest. Alternative aggregation strategies include micro-averaging, which aggregates counts globally across all classes, and weighted averaging, which weights each class by its support.

These metrics, however, have well-known limitations in the SA context. Accuracy is misleading on the imbalanced datasets commonly encountered in real-world applications [40], and standard metrics fail to capture the severity of misclassification, confusing positive with negative is arguably worse than confusing positive with neutral. As a result, modern transformer models often achieve very high and tightly clustered accuracy scores, which limits the discriminative power of SST-2 as a benchmark and motivating next-generation evaluations such as SentiEval [22] that test across diverse task types. Recent work by Herrera-Poyatos et al. [25] highlights an additional concern: the model variability problem (MVP), where LLMs produce different sentiment classifications for identical inputs across runs due to stochastic decoding. This phenomenon challenges the reproducibility of reported benchmark scores and demands new evaluation protocols that account for output variance.

A fundamental limitation specific to sentiment analysis is that ground truth itself is subjective: reasonable annotators may disagree on the sentiment of ambiguous or nuanced text. Inter-annotator agreement metrics such as Krippendorff's α and Cohen's κ therefore serve as a soft upper bound on meaningful model performance. Datasets with low inter-annotator agreement (common in fine-grained and implicit sentiment tasks) constrain the maximum achievable accuracy, making it essential to report agreement statistics alongside model performance. For regression tasks such as multimodal sentiment intensity prediction, Pearson correlation and MAE should be reported jointly, as correlation captures ranking quality while MAE captures calibration. Additionally, the distinction between ordinal and nominal misclassification is often overlooked: standard F1 treats all errors equally, but in ordinal settings (e.g., SST-5), predicting 1-star for a 5-star review is qualitatively worse than predicting 4-star, suggesting that metrics such as mean absolute error or quadratic weighted kappa may be more appropriate.

Taken together, these benchmarks and evaluation protocols have not merely measured progress in sentiment analysis, but actively shaped it. Limitations in early datasets motivated feature-based learning, benchmark saturation exposed the ceilings of classical models, and increasingly complex tasks—such as aspect-based, multimodal, and cross-lingual sentiment analysis—drove the adoption of more expressive architectures. With this empirical context established, we now trace the methodological evolution of sentiment analysis models, highlighting how successive paradigms emerged in response to these evaluation challenges.

The evolution of benchmark datasets has not only measured progress in sentiment analysis but also shaped it. Early datasets such as MR and IMDb emphasized document-level polarity, enabling steady improvements in classification accuracy, while more recent benchmarks introduce fine-grained, multimodal, and cross-lingual challenges that remain far from solved. Notably, the saturation of performance on widely used datasets such as SST-2 limits their discriminative power, making it increasingly difficult to distinguish between modern architectures. At the same time, the subjective nature of sentiment annotation imposes an inherent ceiling on achievable performance, suggesting that future evaluation frameworks must move beyond static accuracy metrics toward robustness, variability, and real-world deployment criteria.

4 Methodological Evolution

This section traces the development of sentiment analysis methods across five major paradigms, progressing from rule-based lexicons to trillion-parameter language models. Table 2 provides a high-level comparative overview, while Table 3 later in this section offers a more granular performance comparison across representative models. Throughout this section, we connect each paradigm back to the formal taxonomy of Section 2, noting how different parameterizations of $P(y \mid \mathbf{x}; \boldsymbol{\theta})$ address different granularity levels, task formulations, and domains with varying degrees of success.

Table 2: Comparative overview of sentiment analysis paradigms.

Paradigm	Era	Repr. Model	IMDb Acc.	SST-2 Acc.	Params	Training
Lexicon	2002+	VADER	~65%–70%	~60%–70%	0	None
Trad. ML	2002+	SVM	~86%–89%	~80%–83%	~10K	Supervised
Deep Learn.	2014+	BiLSTM-Att	~89%–92%	~87%–89%	~5M	Supervised
Transformer	2018+	RoBERTa	~95%–96%	96.4%	355M	Pretrain+FT
LLM	2023+	GPT-4	~93%–96%	~95%–97%	>1T	Zero/Few-shot

Table 3: Detailed performance comparison of representative models across paradigms on standard SA benchmarks. Reported accuracies (%) are from original papers or SentiEval [22].

Model	Year	SST-2	IMDb	Params	Paradigm
VADER [8]	2014	65–70	65–70	–	Lexicon
SVM+BoW [1]	2002	78–82	86–89	~10K	Trad. ML
TextCNN [12]	2014	87–88	90–91	~1M	Deep Learn.
BiLSTM-Att [15]	2016	88–90	90–92	~5M	Deep Learn.
BERT-base [16]	2018	94.9	93–95	110M	Transformer (FT)
RoBERTa [17]	2019	96.4	95–96	355M	Transformer (FT)
XLNet [18]	2019	96–97	95–96	340M	Transformer
ModernBERT [58]	2024	96–97	95–96	395M	Transformer (FT)
GPT-3.5 (0-shot) [22]	2023	88–92	88–91	175B	LLM (Zero-shot)
GPT-4 (0-shot) [22]	2024	94–96	93–95	>1T†	LLM (Zero-shot)
GPT-4 (5-shot) [22]	2024	95–97	94–96	>1T†	LLM (Few-shot)
LLaMA-3-70B (0-shot)	2024	93–95	92–94	70B	LLM (Zero-shot)
Mixtral-8x22B (0-shot)	2024	92–95	92–94	141B*	MoE LLM (Zero-shot)
Gemini 1.5 Pro (0-shot)	2025	94–96	93–95	>500B†	LLM (Zero-shot)

Note: *Total parameters in MoE; active parameters per token are lower. †Estimated model scale based on publicly available information. Recent LLM results are based on prompt-based zero-shot evaluations and are not directly comparable to fine-tuned models.

4.1 Lexicon-Based Methods

Lexicon-based approaches represent the earliest systematic methods for sentiment analysis, relying on pre-compiled dictionaries that associate words or phrases with sentiment scores [6–8,72,93,94]. SentiWordNet [6,7] assigns positivity, negativity, and objectivity scores to each WordNet synset, enabling the computation of aggregate document sentiment as:

$$S(d) = \frac{1}{|d|} \sum_{w \in d} \left(s^{+}(w) - s^{-}(w) \right) \tag{6}$$

where $|d|$ denotes the number of tokens in document d, and $s^{+}(w)$ and $s^{-}(w)$ represent the positive and negative sentiment scores assigned to word w by the underlying lexicon. This formulation assumes that each word contributes independently and equally to the overall sentiment, ignoring contextual effects such as negation, intensification, and compositional semantics. The resulting score $S(d)$ is an aggregate polarity measure whose range depends on the underlying lexicon scores and document length.

VADER (Valence Aware Dictionary and Sentiment Reasoner) [8] extended this approach with rule-based heuristics for handling negation, intensifiers, and punctuation, computing a composite score through:

$$s_{\text{compound}} = \frac{\sum_{i=1}^{n} v_i}{\sqrt{\left(\sum_{i=1}^{n} v_i\right)^2 + \alpha}} \tag{7}$$

where n denotes the number of tokens in the input text for which valence scores are computed, and v_i represents the valence score of the i-th token after applying rule-based adjustments for factors such as negation, degree modifiers, punctuation, and capitalization. The parameter α is a normalization constant (typically set to 15) that controls the scaling of the score and prevents extreme values for large aggregated valence magnitudes. This normalization maps the aggregated valence score to a bounded range in $[-1, 1]$, facilitating consistent comparison across texts of varying lengths.

A third notable tool, LIWC (Linguistic Inquiry and Word Count) [93,94], provides psychologically validated word categories that extend beyond simple polarity to capture cognitive processes, social dynamics, and affective dimensions, and has been widely applied to sentiment analysis in social media and healthcare contexts despite being primarily designed for psychological research.

The strengths of lexicon-based methods lie in their interpretability, domain independence, and zero-resource applicability. However, they fundamentally cannot handle context-dependent sentiment, compositional semantics (e.g., "not bad"), sarcasm, or domain-specific terminology [4,95]. Kennedy and Inkpen [95] demonstrated that contextual valence shifters can partially address negation, but systematic limitations remain, motivating the transition to data-driven approaches.

4.2 Traditional Machine Learning

Machine learning approaches reformulated sentiment analysis as a supervised classification problem, learning discriminative patterns from labeled data [1,9–11].

4.2.1 Support Vector Machines

SVMs became the dominant approach for text classification following Joachims' seminal work [9]. For sentiment analysis, each document is represented as a feature vector $\mathbf{x} \in \mathbb{R}^d$ (typically using bag-of-words or TF-IDF features) and classified by finding the maximum-margin hyperplane. The SVM optimization problem takes the form:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \tag{8}$$

where $\mathbf{w}$ is the weight vector defining the separating hyperplane, b is the bias term, and ξ_i are slack variables that allow for margin violations in non-separable data. The parameter $C > 0$ controls the trade-off between maximizing the margin (via minimizing $\|\mathbf{w}\|^2$) and penalizing classification errors through the slack variables. Here, $y_i \in \{-1, +1\}$ denotes the binary class label associated with input $\mathbf{x}_i$. This formulation corresponds to the soft-margin support vector machine, which allows for misclassification in exchange for improved generalization on non-linearly separable data.

The dual formulation with kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ enables non-linear classification:

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{9}$$

where α_i are the Lagrange multipliers obtained from the dual optimization problem in Eq. (8), and $K(\mathbf{x}_i, \mathbf{x})$ denotes a kernel function that computes an inner product in a (possibly high-dimensional) feature space, enabling non-linear classification. Only a subset of training instances with $\alpha_i > 0$ (the support vectors) contribute to the decision function. The $\text{sgn}(\cdot)$ function returns the sign of its argument, determining the predicted class label.

Pang et al. [1] demonstrated that SVMs with unigram features achieved 82.9% accuracy on their movie review corpus, significantly outperforming lexicon-based methods and establishing machine learning as the dominant paradigm. Subsequent work explored n-gram features, part-of-speech tags [96], and domain adaptation techniques [29].

4.2.2 Naïve Bayes

The Naïve Bayes classifier assumes feature independence and applies Bayes' theorem [10,97]:

$$P(y \mid \mathbf{x}) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(\mathbf{x})} \tag{10}$$

Here, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ represents the feature vector corresponding to the input text, where each x_i typically denotes the presence or frequency of a token. This formulation relies on the Naïve Bayes assumption that features are conditionally independent given the class label, i.e., $P(\mathbf{x} \mid y) = \prod_{i=1}^{n} P(x_i \mid y)$. The denominator $P(\mathbf{x})$ acts as a normalization constant and is typically omitted during prediction, as it is independent of the class label. In practice, classification is performed by selecting the class with the highest posterior probability, i.e., $\hat{y} = \arg\max_y P(y \mid \mathbf{x})$.

Despite the strong independence assumption, Naïve Bayes performs surprisingly well for text classification due to the high dimensionality of the feature space, achieving competitive results with far less training time than SVMs [1,97]. The multinomial variant, which models word counts rather than binary presence, is particularly effective for longer documents [10].

4.3 Deep Learning Approaches

Deep learning methods introduced the ability to learn hierarchical feature representations directly from raw text, eliminating the need for manual feature engineering [2,98,99]. This subsection traces the key architectural developments that shaped the deep learning era of sentiment analysis.

4.3.1 Word Embeddings

The transition from sparse bag-of-words representations to dense, continuous word embeddings marked a foundational shift. Word2Vec [100,101] learns distributed representations through either the skip-gram or continuous bag-of-words (CBOW) objective. The skip-gram model maximizes:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log P(w_{t+j} \mid w_t) \tag{11}$$

where T denotes the total number of tokens in the corpus, w_t is the token at position t, and c is the context window size defining how many neighboring tokens are considered. In practice, the summation over context positions is restricted to valid indices within the sequence boundaries. The conditional probability $P(w_{t+j} \mid w_t)$ is typically parameterized using a softmax function over the vocabulary based on learned word

embeddings. This objective maximizes the likelihood of observing context words given a center word, thereby learning distributed word representations that capture semantic relationships.

GloVe [102] subsequently combined global matrix factorization with local context windows by optimizing a weighted least-squares objective on co-occurrence statistics:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\mathbf{w}_i^\top \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2 \tag{12}$$

where X_{ij} denotes the co-occurrence count of word j in the context of word i within a predefined window over the corpus. Here, $\mathbf{w}_i$ and $\tilde{\mathbf{w}}_j$ are the word and context embedding vectors for words i and j, respectively, and b_i and $\tilde{b}_j$ are their associated bias terms. The function $f(\cdot)$ is a weighting function that down-weights the influence of extremely frequent or rare co-occurrences to improve training stability. This objective enforces that the dot product of word embeddings approximates the logarithm of word co-occurrence counts, thereby capturing global statistical structure of the corpus. These embeddings capture semantic relationships (e.g., "king" – "man" + "woman" ≈ "queen") and provide substantially better input representations for downstream SA models [85,102].

4.3.2 Convolutional Neural Networks

Kim [12] introduced the TextCNN architecture, applying 1D convolutions over word embedding sequences. For an input sentence represented as a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ (where n is the sequence length and d is the embedding dimension), a convolution filter $\mathbf{w} \in \mathbb{R}^{h \times d}$ of window size h produces a feature map:

$$c_i = \text{ReLU}(\mathbf{w} \cdot \mathbf{X}_{i:i+h-1} + b) \tag{13}$$

where $\mathbf{X}_{i:i+h-1} \in \mathbb{R}^{h \times d}$ denotes the submatrix of input embeddings corresponding to a window of h consecutive tokens starting at position i. Here, $\mathbf{w} \in \mathbb{R}^{h \times d}$ is a convolutional filter (kernel) and b is a scalar bias term. The operation $\mathbf{w} \cdot \mathbf{X}_{i:i+h-1}$ denotes the element-wise inner product between the filter and the input window, typically implemented as a flattened dot product. The resulting value c_i represents a scalar feature capturing the presence of a specific pattern in the input at position i. The ReLU$(\cdot)$ activation function introduces non-linearity by mapping negative values to zero.

Max-over-time pooling then extracts the most salient feature: $\hat{c} = \max(c_1, c_2, \ldots, c_{n-h+1})$. Multiple filters with varying window sizes (typically $h \in \{3, 4, 5\}$) capture n-gram patterns at different scales. Kim's TextCNN achieved 88.1% on SST-2, demonstrating that even relatively simple architectures with pre-trained embeddings could yield strong SA performance [12].

4.3.3 Recurrent Neural Networks and LSTM

Recurrent neural networks (RNNs) process sequences token-by-token, maintaining a hidden state that captures sequential dependencies. However, vanilla RNNs suffer from the vanishing gradient problem [103], which limits their ability to learn long-range dependencies. The Long Short-Term Memory (LSTM) architecture [13] addresses this through a gating mechanism that regulates information flow:

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad \text{(forget gate)} \tag{14}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad \text{(input gate)} \tag{15}$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \quad \text{(candidate)} \tag{16}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad \text{(cell state)} \tag{17}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \quad \text{(output gate)} \tag{18}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad \text{(hidden state)} \tag{19}$$

where $\mathbf{x}_t$ is the input vector at time step t, $\mathbf{h}_{t-1}$ is the hidden state from the previous time step, and $\mathbf{c}_t$ denotes the cell state that carries long-term information. The notation $[\mathbf{h}_{t-1}, \mathbf{x}_t]$ denotes the concatenation of the previous hidden state and current input vector, which is linearly transformed by weight matrices $\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_c$, and $\mathbf{W}_o$ with corresponding biases. The sigmoid function $\sigma(\cdot)$ maps values to the range $[0, 1]$ and is used for gating, while $\tanh(\cdot)$ produces values in $[-1, 1]$ and is used to generate candidate cell states. The operator $\odot$ denotes element-wise multiplication. The forget gate $\mathbf{f}_t$ controls which information from the previous cell state is retained, the input gate $\mathbf{i}_t$ determines how much new information is incorporated, and the output gate $\mathbf{o}_t$ regulates the information exposed to the hidden state. The cell state $\mathbf{c}_t$ serves as a memory mechanism that enables the network to preserve information over long time horizons, mitigating the vanishing gradient problem in standard recurrent networks.

Tai et al. [104] extended LSTMs to tree-structured topologies (Tree-LSTM) that operate over parse trees, achieving 51.0% accuracy on the fine-grained SST-5 task. Bidirectional LSTMs (BiLSTMs) process sequences in both directions, and when combined with attention mechanisms [15], they achieved state-of-the-art performance in the pre-transformer era. More recently, Nkhata et al. [105] demonstrated that fine-tuning BERT with a bidirectional LSTM layer produces further improvements for fine-grained movie review sentiment classification, suggesting that recurrent layers can still complement transformer representations. Fig. 6 illustrates the structural differences between vanilla RNNs and gated recurrent architectures that motivated their adoption in early sentiment analysis models.

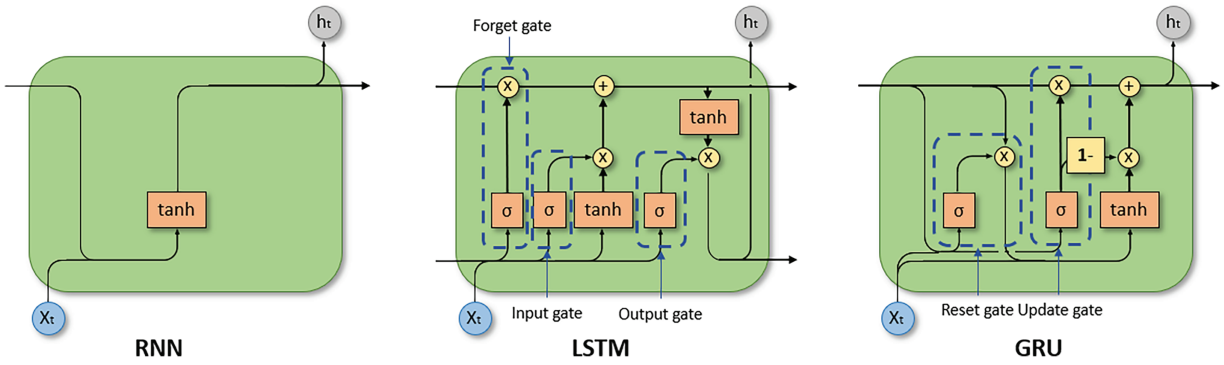

Figure 6: Comparison of recurrent neural network architectures. Vanilla RNNs propagate information through a single hidden state, while gated variants such as LSTM and GRU introduce explicit control mechanisms to regulate information flow and mitigate the vanishing gradient problem. These architectures formed the backbone of pre-transformer deep learning approaches to sentiment analysis. Adapted from [106].

4.3.4 Attention Mechanisms

The attention mechanism, introduced by Bahdanau et al. [14], allows models to selectively focus on relevant parts of the input. For sentiment analysis, attention-based models compute a weighted sum of hidden states:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{n} \exp(e_k)}, \quad \mathbf{v} = \sum_{t=1}^{n} \alpha_t \mathbf{h}_t \tag{20}$$

where $e_t = \mathbf{u}^\top \tanh(\mathbf{W}_h \mathbf{h}_t + \mathbf{b})$ is the attention energy. $\mathbf{h}_t$ denotes the hidden representation of the input at position t, typically obtained from a recurrent or encoder-based model. The scalar e_t represents the

unnormalized relevance score of the t-th input with respect to the attention mechanism, computed using learnable parameters $\mathbf{W}_h$, $\mathbf{u}$, and bias $\mathbf{b}$. The coefficients α_t are attention weights obtained via a softmax function, ensuring that $\sum_{t=1}^{n} \alpha_t = 1$, and quantify the relative importance of each input position. The vector $\mathbf{v}$ is a context vector computed as a weighted sum of hidden representations, aggregating relevant information across the sequence.

Wang et al. [15] applied aspect-specific attention to LSTM hidden states for ABSA, conditioning the attention weights on the aspect embedding. Yang et al. [107] introduced hierarchical attention networks (HAN) with word-level and sentence-level attention for document classification, naturally capturing the multi-granularity structure of sentiment in long documents.

Despite their success, deep learning models based on CNNs and RNNs remained fundamentally limited by their reliance on task-specific training and their inability to fully exploit large-scale unlabeled text. These constraints motivated a shift toward pre-trained language models that could acquire general linguistic knowledge once and transfer it efficiently across sentiment analysis tasks.

4.4 Pre-Trained Transformers

The transformer architecture [108] and its pre-trained variants fundamentally transformed sentiment analysis by introducing large-scale transfer learning, enabling models pre-trained on massive unlabeled corpora to be adapted to sentiment tasks with relatively small amounts of labeled data.

4.4.1 The Transformer Architecture

The transformer [108] replaces recurrence with *self-attention*, computing pairwise interactions between all tokens in parallel. Given an input representation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, the self-attention mechanism computes contextualized representations by projecting $\mathbf{X}$ into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices and applying scaled dot-product attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{21}$$

In the input representation matrix denoted by $\mathbf{X} \in \mathbb{R}^{n \times d}$, n is the sequence length and d the embedding dimension. The queries, keys, and values are obtained as learned linear projections of $\mathbf{X}$, i.e., $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_V$, where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are trainable weight matrices. Typically, $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, where d_k and d_v denote the dimensionalities of the query/key and value vectors, respectively. The matrix $\mathbf{Q}\mathbf{K}^\top$ computes pairwise similarity scores between queries and keys, and the softmax function is applied row-wise to normalize these scores across all key positions for each query, ensuring that attention weights sum to one. The scaling factor $\sqrt{d_k}$ prevents the dot products from becoming excessively large, improving numerical stability and gradient behavior. The resulting output is a weighted combination of value vectors, producing context-aware representations for each input position. Fig. 7 visualizes the scaled dot-product attention operation alongside its multi-head extension, showing how query, key, and value projections are combined to produce contextualized token representations.

Multi-head attention extends this by running h parallel attention operations with different learned projections:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{22}$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$. Here, h denotes the number of attention heads, and $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ are learnable projection matrices for the i-th head. Each head operates on lower-dimensional

subspaces, typically of size d_k/h and d_v/h. The outputs of all heads are concatenated along the feature dimension using Concat$(\cdot)$ and then linearly transformed by a learnable projection matrix $\mathbf{W}_O$ to produce the final output. This multi-head structure enables the model to capture diverse relationships by attending to information from different representation subspaces. Each head can attend to different aspects of the input, for sentiment analysis, some heads may focus on opinion words while others attend to negation or aspect terms. The overall encoder stack that forms the backbone of BERT and its variants is depicted in Fig. 8: each encoder block comprises multi-head self-attention followed by a position-wise feed-forward network, with residual connections and layer normalization applied at both sub-layers; BERT-base stacks 12 such blocks, while BERT-large extends this to 24, providing the representational capacity that underpins the performance gains reported on SST-2 and IMDb.

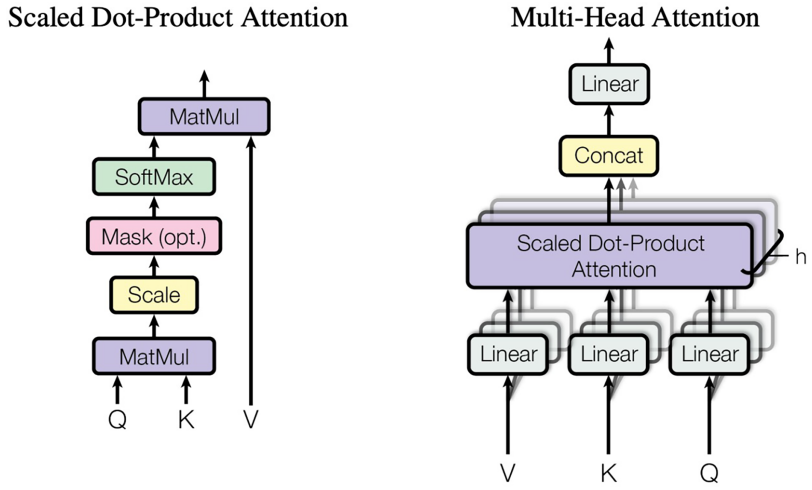

Figure 7: Scaled dot-product self-attention mechanism [108]. Input token embeddings are linearly projected into query, key, and value representations. Attention weights are computed via normalized dot products between queries and keys, enabling each token to selectively aggregate information from all other tokens in the sequence. Adapted from [109].

4.4.2 BERT and Variants

BERT (Bidirectional Encoder Representations from Transformers) [16] pre-trains a deep transformer encoder using masked language modeling (MLM) and next sentence prediction (NSP) on large unlabeled corpora:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid \mathbf{x}_{\setminus \mathcal{M}}; \boldsymbol{\theta}) \tag{23}$$

where $\boldsymbol{\theta}$ is the parameters of the transformer model (e.g., BERT) learned during pre-training. $\mathcal{M}$ denotes the set of masked token positions in the input sequence. The input $\mathbf{x}$ is first corrupted by replacing tokens at positions in $\mathcal{M}$ with a special [MASK] token (or, with small probability, random or unchanged tokens), and the model is trained to predict the original tokens x_i at those positions. Accordingly, $P(x_i \mid \mathbf{x}_{\setminus \mathcal{M}}; \boldsymbol{\theta})$ denotes the conditional probability of the original token given the corrupted input sequence, where masked positions remain present but altered rather than removed. This objective can be interpreted as maximizing the log-likelihood of masked tokens under a corruption distribution over input sequences.

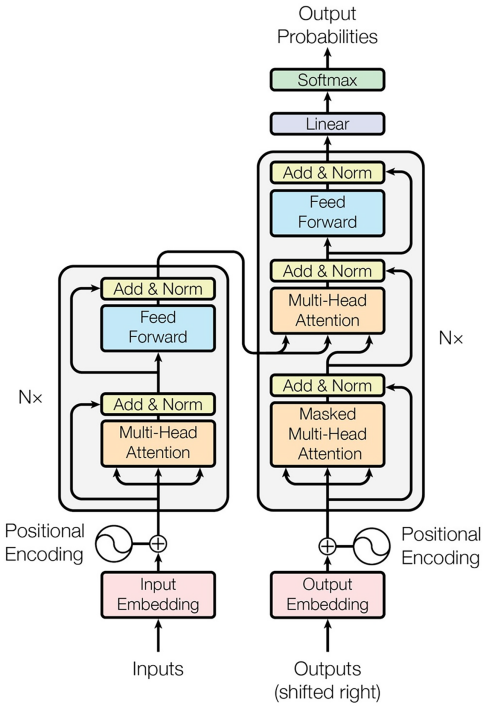

Figure 8: Transformer encoder architecture and BERT scaling [108]. Each encoder block consists of multi-head self-attention followed by position-wise feed-forward layers with residual connections and layer normalization. BERT-base stacks 12 such layers, while BERT-large extends this depth to 24 layers, enabling greater representational capacity for downstream sentiment analysis tasks. Adapted from [110].

BERT-base (110M parameters) achieved 94.9% on SST-2 upon release, representing a dramatic improvement over prior methods [16]. Several subsequent variants optimized different aspects of the pre-training and architecture. RoBERTa [17] removed NSP, trained with larger batches and more data, and used dynamic masking, pushing SST-2 accuracy to 96.4%. XLNet [18] combined autoregressive modeling with permutation-based training to capture bidirectional context without masking artifacts. DistilBERT [111] applied knowledge distillation to compress BERT to 66M parameters (40% smaller) while retaining 97% of its performance, enabling deployment in resource-constrained settings.

For domain-specific applications, Sun et al. [19] established best practices for fine-tuning BERT on sentiment tasks, finding that gradual unfreezing and discriminative learning rates significantly improve convergence. Bello et al. [112] and Batra et al. [113] demonstrated BERT's effectiveness for tweet and software review sentiment analysis, respectively, while Penha and Hauff [114] probed BERT's internal representations to understand what it learns about domain-specific sentiment. InstructABSA [80] takes a different tack, reformulating ABSA as an instruction-following task for instruction-tuned transformers and achieving competitive results with minimal task-specific fine-tuning.

4.4.3 ModernBERT

Warner et al. [58] introduced ModernBERT in December 2024 as a modernized encoder architecture that incorporates long-context training, rotary positional embeddings (RoPE), and optimized attention kernels. The key innovations include an extended context length of 8192 tokens (compared to BERT's 512), enabling processing of full-length reviews without truncation; training on 2 trillion tokens, an order of magnitude more than the original BERT; rotary positional embeddings (RoPE) replacing absolute position

encodings to improve length generalization; and Flash Attention with unpadding for efficient inference. ModernBERT outperforms BERT, RoBERTa, and DeBERTa across classification tasks while maintaining efficient inference, making it a strong candidate for production SA systems that require the fine-tuned model paradigm. Recent work combining ModernBERT with SHAP and LIME has further demonstrated enhanced explainability for sentiment classification [68].

4.5 Large Language Models

The emergence of large language models has introduced a qualitatively new paradigm in which models can perform sentiment analysis without task-specific fine-tuning, relying instead on zero-shot prompting, few-shot in-context learning, and chain-of-thought reasoning.

4.5.1 Zero-Shot and Few-Shot Sentiment Analysis

LLMs leverage their pre-training on massive corpora to perform SA through natural language instructions. In zero-shot mode, the model receives only a task description such as "*Classify the sentiment of the following review as positive or negative: [review text]*." In few-shot mode, k labeled examples are prepended as context [20,115]:

$$P(y \mid \mathbf{x}, \mathcal{D}_k; \boldsymbol{\theta}) = P\big(y \mid [\mathbf{x}_1, y_1, \ldots, \mathbf{x}_k, y_k, \mathbf{x}]; \boldsymbol{\theta}\big) \tag{24}$$

where $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{k}$ denotes the set of demonstration examples, and $[\mathbf{x}_1, y_1, \ldots, \mathbf{x}_k, y_k, \mathbf{x}]$ represents the concatenation of input-output pairs into a single prompt sequence. The probability is computed by a pre-trained language model with fixed parameters $\boldsymbol{\theta}$ via autoregressive token prediction.

Zhang et al. [22] conducted the most comprehensive evaluation to date using SentiEval, testing across 13 tasks and 26 datasets. Their findings reveal a nuanced picture: in zero-shot settings, LLMs such as ChatGPT and GPT-4 achieve satisfactory performance on simple binary classification but lag behind fine-tuned SLMs on complex tasks such as ABSA and fine-grained classification; in few-shot settings, LLMs significantly outperform SLMs, suggesting particular value when annotation resources are limited; and prompt design has a substantial impact, with performance variance across five different prompts reaching up to 15 percentage points on some datasets. To summarize these trends visually, Fig. 9 presents an illustrative, survey-level comparison of zero-shot and fine-tuned sentiment analysis performance across representative model paradigms.

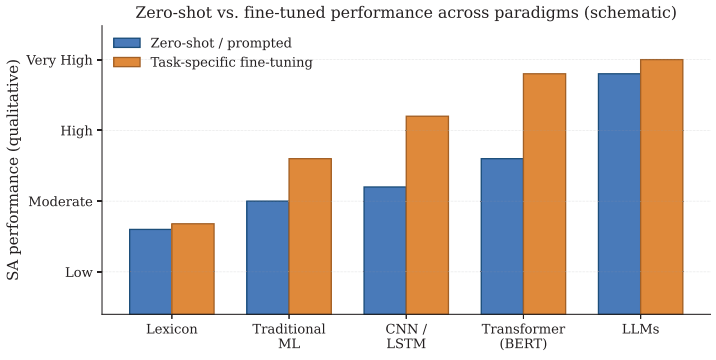

Figure 9: Illustrative comparison of zero-shot (prompted) and task-specific fine-tuned relative performance (%) across major sentiment analysis model paradigms. Bar heights reflect normalized trends consistently reported across multiple benchmarks and studies, rather than absolute accuracy values from a single dataset or experimental configuration (This figure was generated with the assistance of OpenAI GPT 5.2).

Rathje et al. [24] evaluated GPT-3.5, GPT-4, and GPT-4-Turbo on 15 psychological text analysis datasets across 12 languages, finding correlations in the range $r = 0.59–0.77$ with human annotations, compared to $r = 0.20–0.30$ for dictionary-based methods, a dramatic improvement for multilingual SA. Krugmann and Hartmann [23] provided a systematic analysis of generative AI's potential for SA, identifying prompt engineering, cost efficiency, and hallucination risk as key considerations for practical deployment.

4.5.2 Open-Source vs. Closed-Source LLMs

The LLM landscape for sentiment analysis is bifurcated between proprietary models and open-source alternatives. On the open-source side, Meta's Llama-3 [21] family culminated in the Llama-3.1-405B model, which Meta reports was pretrained on approximately 15 trillion tokens and employs grouped query attention (GQA) for efficient inference; it demonstrates strong performance on SA benchmarks, particularly when fine-tuned with domain-specific data. On the proprietary side, GPT-4 and GPT-4o represent the current state of the art; Shen and Zhang [52] showed that GPT-4o with few-shot prompting matches FinBERT's fine-tuned performance on financial news sentiment, while Bhatia et al. [87] demonstrated that FinTral, a multimodal financial LLM family built on Llama 3, achieves GPT-4-level performance on financial SA tasks.

A comprehensive comparison by Zhu et al. [62] in their Model Arena for Cross-lingual Sentiment Analysis evaluated XLM-R, Llama-3, and GPT-4, finding that GPT-4 maintains a consistent advantage in zero-shot cross-lingual settings, but the gap narrows substantially when open-source models are fine-tuned on target-language data. The most recent human-vs.-LLM comparison study [53] tested 33 human annotators against 8 LLM variants (GPT-3.5, GPT-4, GPT-4o, Gemini, Llama-3.1, Mixtral) on 100 items, finding that GPT-4o achieved the highest agreement with expert consensus.

4.5.3 Chain-of-Thought and Reasoning-Augmented Sentiment Analysis

A particularly promising development is the use of chain-of-thought (CoT) prompting [116] to improve LLM performance on sentiment tasks that require implicit reasoning. Fei et al. [54] introduced **THOR** (Three-Hop Reasoning), a structured CoT framework specifically designed for implicit sentiment analysis. The three hops consist of identifying the stimulating aspect, inferring the opinion holder's emotional reaction, and deriving the sentiment polarity. Formally, the reasoning chain augments the prediction:

$$P(y \mid \mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathcal{R}} P(y \mid \mathbf{x}, \mathcal{R}; \boldsymbol{\theta}) \cdot P(\mathcal{R} \mid \mathbf{x}; \boldsymbol{\theta}) \tag{25}$$

where $\mathcal{R}$ denotes a reasoning chain (e.g., a sequence of intermediate reasoning steps generated via chain-of-thought prompting). This formulation interprets reasoning as a latent variable that mediates the prediction. In practice, however, large language models approximate this marginalization by generating one or a small number of reasoning paths rather than explicitly summing over all possible $\mathcal{R}$.

THOR demonstrated significant improvements on implicit sentiment datasets where direct classification fails. Duan and Wang [55] proposed a complementary framework called **SAoT** (Sentiment Analysis of Thought) using ERNIE-Bot-4, which decomposes implicit sentiment analysis into structured reasoning steps with explicit intermediate outputs, achieving improvements over direct prompting approaches.

However, Zheng et al. [56] provide an important critical reassessment, demonstrating that CoT does not uniformly improve SA performance across all task types. Their analysis suggests that CoT is most beneficial for tasks requiring pragmatic inference (implicit sentiment, sarcasm) but can actually hurt performance on straightforward polarity classification, where the reasoning overhead introduces unnecessary error propagation. Recent extensions include multi-chain CoT for ABSA [57], which employs multiple parallel reasoning

chains to handle the complexity of aspect-level sentiment, and graph-enhanced reasoning [117], which combines graph neural networks with prompt-based reasoning for implicit aspect-based sentiment analysis.

4.5.4 Model Variability Problem

Herrera-Poyatos et al. [25] identified the *model variability problem* (MVP) as a fundamental challenge for LLM-based sentiment analysis. MVP manifests as inconsistent sentiment classifications for identical inputs across multiple inference runs, arising from three principal sources. First, stochastic decoding with temperature-based sampling ($T > 0$) introduces randomness into token selection, where the probability of generating token w is modulated as:

$$P(w \mid \mathbf{w}_{<t}; \boldsymbol{\theta}) = \frac{\exp(z_w/T)}{\sum_{w'} \exp(z_{w'}/T)} \tag{26}$$

where z_w denotes the logit corresponding to token w produced by the model given the previous context $\mathbf{w}_{<t}$, and $T > 0$ is the temperature parameter controlling the sharpness of the distribution. Lower values of T concentrate probability mass on high-scoring tokens, while higher values increase output diversity. Second, prompt sensitivity means that minor rephrasing of prompts can shift predictions across decision boundaries. Third, systematic biases in pre-training corpora propagate into inconsistent sentiment judgments. The authors emphasize the role of explainability frameworks in mitigating MVP, connecting to broader XAI efforts discussed in Section 6.5.

To quantify MVP in a standardized manner, several measurement approaches can be defined. The *classification instability rate* (CIR) measures the proportion of inputs that receive different sentiment labels across k repeated inference runs:

$$\text{CIR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\exists\, j, l \in \{1, \ldots, k\} : \hat{y}_i^{(j)} \neq \hat{y}_i^{(l)}\right] \tag{27}$$

where $\hat{y}_i^{(j)}$ denotes the predicted label for input i on run j. $\mathbf{1}[\cdot]$ denotes the indicator function, which equals 1 if its condition is true and 0 otherwise, and k is the number of repeated inference runs for the same input. Thus, CIR $\in [0, 1]$, with CIR = 0 indicating perfectly stable classifications across runs and CIR = 1 indicating that every input receives at least one inconsistent label across the k runs. Complementarily, entropy-based variability captures the spread of output distributions: for each input $\mathbf{x}_i$, the empirical entropy $H_i = -\sum_{c \in \mathcal{C}} \hat{p}_c \log \hat{p}_c$ over classes, averaged across the dataset, quantifies overall model uncertainty. The SentiEval evaluation [22] provides indirect evidence of prompt-induced variability, documenting 10%–15% performance swings across five different prompt formulations on the same dataset.

It is important to distinguish between *aleatoric* variability (inherent in stochastic decoding at $T > 0$, reducible by setting $T \to 0$ or using greedy decoding) and *epistemic* variability (arising from prompt sensitivity and training data biases, which persists even at $T = 0$). Current mitigation strategies ensemble averaging over multiple runs, temperature annealing, majority voting are empirically motivated but lack formal convergence guarantees. Developing principled methods to bound and reduce MVP remains an important open problem with direct implications for deployment in regulated industries such as finance and healthcare.

4.5.5 Prompt Engineering for Sentiment Analysis

Effective prompt design is critical for LLM-based SA. Stilwell and Inkpen [118] show that prompt wording significantly affects model performance on the IMDb sentiment benchmark and that prompt

learning can substantially improve classification accuracy. They also investigate explainability methods for prompt-based classifiers, finding that explanations generated directly by LLMs are judged by humans to be more adequate and trustworthy than traditional XAI approaches such as LIME or SHAP. The prompts evaluated in their study include instructional, completion, and question-based formulations that guide the model to produce binary sentiment labels. Gu et al. [119] introduced prompt tuning as an alternative to manual prompt engineering, learning continuous prompt vectors that are prepended to the input and achieving few-shot performance comparable to full fine-tuning. More systematically, prompting strategies for SA can be organized into a hierarchy of increasing sophistication: (1) *zero-shot instruction prompts*, which provide only the task description and output format which are effective for simple binary classification but insufficient for nuanced tasks; (2) *few-shot demonstration prompts*, which prepend labeled examples, particularly effective when demonstrations are drawn from the target domain and cover diverse sentiment patterns; (3) *chain-of-thought prompts*, which request step-by-step reasoning before classification, most beneficial for implicit sentiment and sarcasm, but potentially harmful for straightforward polarity tasks due to error propagation [56]; and (4) *role-specification prompts*. Fig. 10 illustrates how variations in prompt formulation can lead to different sentiment interpretations for the same input text, even when using the same underlying language model.

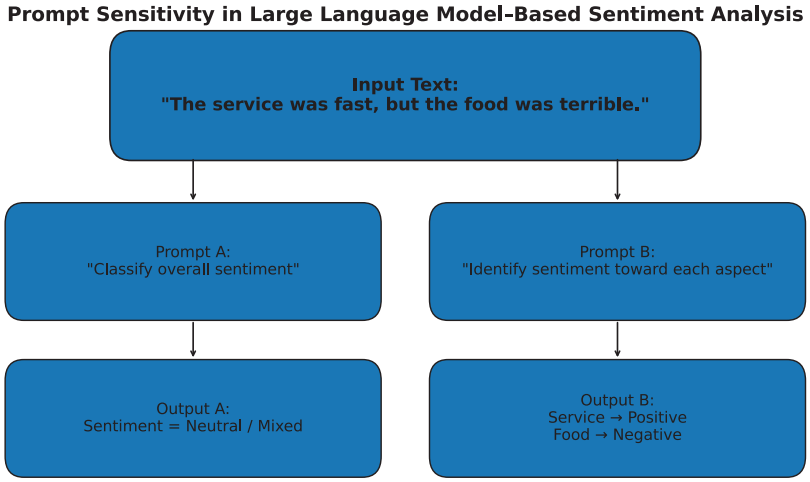

Figure 10: Illustration of prompt sensitivity in large language model–based sentiment analysis. The same input text can yield different sentiment interpretations depending on prompt formulation, highlighting the dependence of LLM-based sentiment outputs on task framing and instruction design (This figure was generated with the assistance of OpenAI GPT 5.2).

4.5.6 Ensemble Approaches

A comprehensive survey of ensemble strategies for LLMs [71] catalogs methods for combining multiple LLM outputs to improve SA robustness, including majority voting across model variants, weighted aggregation based on confidence scores, and mixture-of-experts architectures that route inputs to specialized sub-models. Ensemble approaches directly address the MVP by averaging over stochastic outputs, though at increased computational cost.

4.6 Multimodal Sentiment Analysis

Human communication is inherently multimodal, conveying sentiment through not only words but also tone of voice, facial expressions, and gestures [34]. Multimodal sentiment analysis (MSA) integrates these

complementary signals, and the choice of fusion strategy, that is, how modality-specific representations are combined, is a central design decision.

4.6.1 Fusion Strategies

Given modality-specific representations $\mathbf{h}_t$ (text), $\mathbf{h}_a$ (audio), and $\mathbf{h}_v$ (visual), fusion strategies determine how these are combined [34,35,91].

Early fusion concatenates raw or lightly processed features before a joint model:

$$\mathbf{h}_{\text{fused}} = [\mathbf{h}_t; \mathbf{h}_a; \mathbf{h}_v] \tag{28}$$

where, $[\cdot;\cdot;\cdot]$ denotes vector concatenation along the feature dimension.

Late fusion trains separate modality-specific classifiers and combines their predictions:

$$\hat{y} = g(f_t(\mathbf{h}_t), f_a(\mathbf{h}_a), f_v(\mathbf{h}_v)) \tag{29}$$

where g is a combination function (e.g., weighted average, learned MLP). f_t, f_a, and f_v are modality-specific prediction functions (e.g., classifiers), and g is a fusion function that combines their outputs.

Tensor fusion [35] computes the outer product across modalities to capture multiplicative interactions:

$$\mathbf{z} = \begin{bmatrix} \mathbf{h}_t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{h}_a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{h}_v \\ 1 \end{bmatrix} \tag{30}$$

The appended 1 captures unimodal and bimodal interactions as subsets of the full tensor. Here, each modality vector is augmented with a constant 1, and $\otimes$ denotes the outer (tensor) product, enabling the model to capture unimodal, bimodal, and trimodal interactions. While tensor fusion is theoretically expressive, it suffers from exponential dimensionality growth.

Cross-modal attention [90] learns which elements of one modality are most relevant to another:

$$\mathbf{h}_{t \leftarrow a} = \text{Attention}(\mathbf{Q}_t, \mathbf{K}_a, \mathbf{V}_a) \tag{31}$$

Here, $\mathbf{Q}_t$, $\mathbf{K}_a$, and $\mathbf{V}_a$ denote query, key, and value projections derived from the text and audio modalities, respectively.

Multi-attention recurrent networks [90] apply this bidirectionally across all modality pairs. More recently, Cai et al. [120] proposed a multi-layer feature fusion approach combined with multi-task learning, demonstrating improvements on CMU-MOSI and CMU-MOSEI, while Ren [121] combined BERT for text encoding with ResNet for image features, achieving 74.5% accuracy on the MAVA-single multimodal benchmark through an attention-based fusion mechanism. Fig. 11 summarizes the principal multimodal fusion paradigms used in sentiment analysis, highlighting the trade-offs between representational expressiveness and computational cost.

4.6.2 Multimodal Transformers

Transformer architectures have been adapted for multimodal inputs, with MMBERT [122] extending BERT's architecture with cross-modal attention layers that fuse text and visual representations. More recently, the explosive growth of multimodal LLMs (MLLMs) has opened new possibilities for sentiment analysis. Yang et al. [36] provide a comprehensive survey of how LLMs are being integrated into text-centric multimodal SA, identifying key challenges in grounding visual and acoustic features within LLM

representations. da Silva et al. [59] introduce MLLMsent, a framework for evaluating whether multimodal LLMs can "see" sentiment in images and text, achieving state-of-the-art results that significantly outperform prior specialist models; their findings suggest that MLLMs' ability to jointly reason over visual and textual modalities provides a natural advantage for tasks that require understanding both content and presentation. A systematic review [60], analyzing 116 multimodal SA studies published between 2018 and 2025, identifies three key trends: the shift from handcrafted fusion to end-to-end learned fusion, the growing dominance of transformer-based architectures, and the persistent challenge of temporal alignment between modalities.

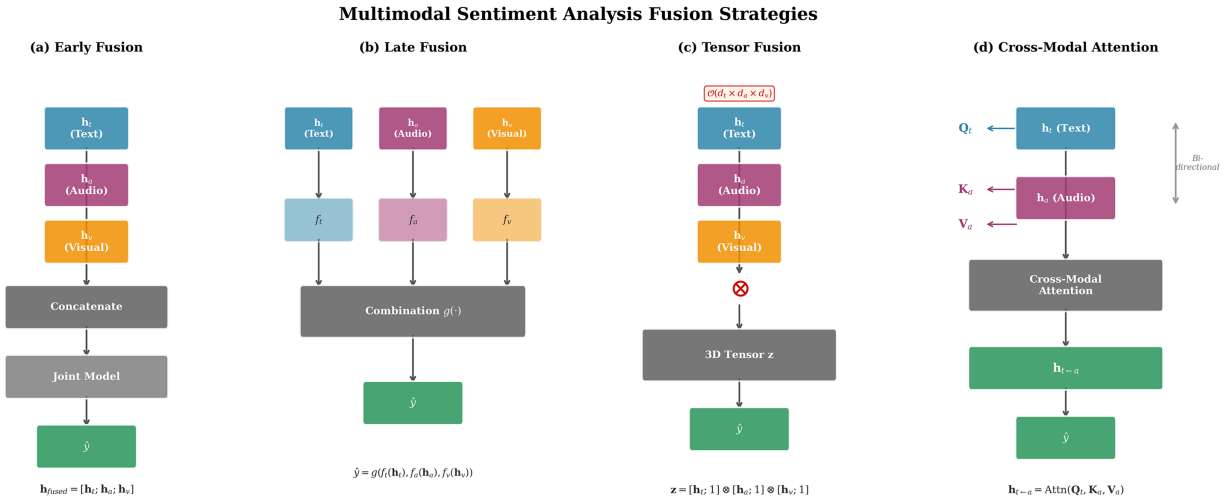

Figure 11: Taxonomy of multimodal fusion strategies for sentiment analysis. (**a**) Early fusion concatenates modality representations before joint modeling. (**b**) Late fusion trains independent classifiers per modality and combines predictions. (**c**) Tensor fusion computes outer products across modalities at exponential dimensionality cost. (**d**) Cross-modal attention enables selective information flow between modality pairs through query–key–value projections (This figure was generated with the assistance of OpenAI GPT 5.2).

A critical architectural distinction in recent multimodal SA is between *early-fusion multimodal transformers*, which tokenize and embed all modalities into a shared sequence space from the initial layers (enabling deep cross-modal interaction but requiring modality-specific tokenizers), and *late-integration approaches*, which employ modality-specific encoders followed by cross-modal attention in upper layers (preserving modality-specific representations but potentially limiting deep interaction). The MLLMsent results [59] suggest that general-purpose multimodal LLMs, which typically use the late-integration paradigm, can leverage their massive pre-training to overcome the interaction depth limitation, achieving competitive results without task-specific architectural design. However, the systematic review [60] notes that text features continue to dominate in most fusion architectures, raising the question of whether current models truly leverage multimodal information or primarily rely on text with minor acoustic/visual augmentation.

4.7 Comparative Analysis across Paradigms

Table 3 provides a detailed comparison across representative models from each paradigm on standard benchmarks. Several patterns emerge from this comparison. First, there is a consistent upward trajectory in performance, with each paradigm shift yielding measurable improvements. Second, the gap between fine-tuned transformers and zero-shot LLMs is surprisingly small on binary classification tasks but widens substantially for fine-grained and aspect-based tasks [22]. Third, the computational cost increases by orders

of magnitude across paradigms: from negligible for lexicon methods to billions of FLOPs for LLM inference, creating a trade-off where practitioners must balance accuracy against computational budget [123].

The paradox of the current era is that LLMs are simultaneously the most capable and the most unreliable SA systems: they achieve near-human performance on many tasks [53] while suffering from the model variability problem [25], prompt sensitivity [22], and occasional hallucination [23]. This makes the choice between fine-tuned SLMs and prompted LLMs a non-trivial engineering decision that depends heavily on the specific deployment context. Fig. 12 visualizes this performance–cost trade-off across representative sentiment analysis model paradigms. it is worth mentioning that Fig. 12 is a survey-level synthesis based on relative trends reported across multiple benchmarks and studies, rather than absolute results from a single dataset or experimental setup.

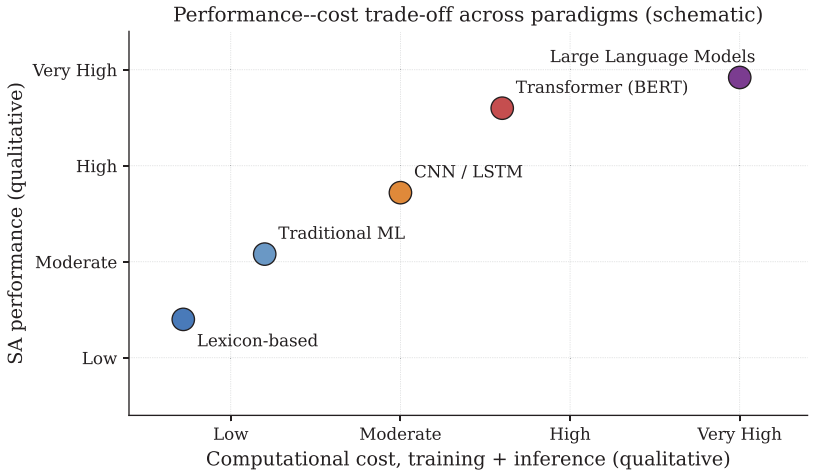

Figure 12: Illustrative performance–cost trade-offs across major sentiment analysis model paradigms. Relative sentiment analysis performance (%) is shown against normalized computational cost (arbitrary units) for training and inference, summarizing trends reported across multiple benchmarks and studies rather than absolute results from a single experimental setup (This figure was generated with the assistance of OpenAI GPT 5.2).

While successive methodological paradigms have delivered substantial performance gains on standard benchmarks, these advances have also revealed new limitations that become evident in real-world deployment. High accuracy on curated datasets does not guarantee robustness to sarcasm, domain drift, cultural variation, or adversarial content. The next section examines these domain-specific challenges in detail, highlighting persistent failure modes that remain unresolved even in the era of large language models.

4.7.1 Failure Modes and Applicability Boundaries

A deeper analysis reveals paradigm-specific failure modes that determine applicability boundaries. Lexicon-based methods fail systematically on compositional semantics ("not bad" → misclassified as negative), figurative language, and any domain where the lexicon lacks coverage: they are best suited for quick, interpretable baselines in well-understood domains. Classical ML methods (SVM, Naïve Bayes) are brittle to out-of-vocabulary terms and domain shift, as their bag-of-words representations cannot generalize beyond the training vocabulary; they remain appropriate for resource-constrained settings with stable, in-domain data. Deep learning models (CNN, LSTM) degrade on short texts where sequential patterns are sparse and on aspect-level tasks without sufficient aspect-annotated training data; they offer the best accuracy-to-cost ratio for medium-scale supervised settings. Fine-tuned transformers (BERT, ModernBERT) achieve near-ceiling

performance on standard benchmarks but struggle with implicit sentiment and sarcasm (where surface-level patterns are misleading) and require per-task fine-tuning that scales poorly across domains; they are the current best choice for production systems with well-defined task boundaries. LLMs excel at zero-shot generalization and few-shot adaptation but suffer from the MVP, prompt sensitivity, hallucinated sentiment labels (particularly for ambiguous inputs), and high inference cost; they are most valuable for low-resource, multi-domain, or rapidly evolving deployment contexts where fine-tuning is impractical.

4.7.2 Practitioner Decision Framework

The choice among SA paradigms involves navigating a multi-dimensional trade-off space. Based on the empirical patterns documented throughout this survey, we propose the following structured decision framework for practitioners:

Task complexity: For binary classification on well-defined domains, fine-tuned transformers (BERT/ModernBERT) offer the best accuracy-to-cost ratio. For ABSA and fine-grained tasks, fine-tuned SLMs outperform zero-shot LLMs [22]. For implicit sentiment and tasks requiring pragmatic inference, CoT-prompted LLMs show advantages [54,55].

Available labeled data: With >1K labeled examples per class, fine-tuning is generally superior. With 5–50 examples, few-shot LLMs become competitive. With zero labeled examples, LLM zero-shot is the only viable option among neural methods, though VADER remains a useful baseline for social media text.

Latency and throughput: Lexicon methods process thousands of documents per second. Inference throughput for transformer classifiers such as BERT varies substantially depending on hardware configuration, sequence length, and batching strategy. Reported performance in practical deployments ranges from tens to thousands of documents per second on modern GPUs. Inference latency for LLM-based sentiment analysis depends strongly on model size, provider infrastructure, and prompt length. Reported response times for API-based systems typically range from sub-second to several seconds per request.

Consistency requirements: For regulated industries (finance, healthcare) where reproducibility is mandated, the MVP makes raw LLM deployment problematic. Ensemble methods [71] or deterministic decoding ($T = 0$) can partially mitigate this, but fine-tuned SLMs with fixed weights offer inherently deterministic inference.

Computational budget: Fine-tuning compact transformer models such as BERT-base can often be completed within a few GPU-hours for moderate-sized datasets, although training time varies significantly with dataset scale, model configuration, and hardware. Token-based pricing for commercial LLM APIs varies widely across providers and model classes, typically ranging from fractions of a cent to several cents per thousand tokens depending on model capability and deployment tier. The Green AI framework [123] advocates reporting computational costs alongside accuracy to enable informed paradigm selection.

A consistent pattern emerges across the methodological evolution of sentiment analysis: performance improvements on standard benchmarks such as SST-2 and IMDb have largely saturated with transformer-based models, while more complex tasks continue to expose fundamental limitations. In particular, fine-grained sentiment classification, aspect-based sentiment analysis, and cross-lingual transfer reveal a widening gap between model architectures. While large language models demonstrate strong zero-shot capabilities, they remain less reliable than fine-tuned transformer models in structured tasks, especially under domain shift. This suggests that recent progress is no longer driven primarily by raw model capacity, but by the ability to balance generalization, stability, and task-specific adaptation.

4.7.3 Synthesis of Paradigm Trade-Offs

While prior comparisons (Tables 2 and 3) emphasize benchmark performance, a broader pattern emerges when sentiment analysis paradigms are evaluated across qualitative capability dimensions. Specifically, recent progress is characterized less by absolute accuracy gains and more by trade-offs between contextual understanding, generalization ability, interpretability, and stability.

Table 4 reveals that the evolution of sentiment analysis has shifted from improving accuracy to navigating trade-offs between generalization, interpretability, and robustness. In particular, while LLMs offer strong zero-shot capabilities, their instability and lack of interpretability present significant challenges for deployment, whereas transformer-based models remain more reliable in controlled settings.

Table 4: Analytical comparison of sentiment analysis paradigms across capability dimensions. Unlike performance-focused comparisons, this table highlights structural trade-offs between paradigms.

Paradigm	Context Sensitivity	Generalization	Interpretability	Stability
Lexicon-based	Low	High (domain-agnostic)	High	High
Traditional ML	Low–Moderate	Moderate	Moderate	High
Deep Learning	Moderate–High	Moderate	Low	High
Transformers	High	High (fine-tuned)	Low–Moderate	High
LLMs	Very High	Very High (zero-shot)	Low	Low (MVP)

5 Domain-Specific Challenges

While the methodological advances described in the previous section have driven substantial performance gains on standard benchmarks, real-world deployment of sentiment analysis systems reveals a constellation of domain-specific challenges that remain only partially addressed. This section examines the most pressing of these challenges. Fig. 13 provides a high-level synthesis of the most commonly reported error categories in sentiment analysis systems, based on recurring patterns observed across benchmark evaluations and qualitative error analyses in the literature.

5.1 Sarcasm and Irony Detection

Sarcasm, where the intended meaning is opposite to the literal meaning, remains one of the most persistent challenges in SA [26,27,124,125]. The utterance "Oh great, another meeting" conveys negative sentiment despite containing the positive word "great," illustrating the fundamental difficulty that sarcasm poses for any system that relies on surface-level lexical cues. Joshi et al. [26] provide a comprehensive survey of automatic sarcasm detection, identifying three broad approaches: rule-based, statistical, and deep learning.

Recent work has specifically evaluated LLM capabilities in this area. Zhang et al. [28] introduced SarcasmBench, the first comprehensive benchmark for evaluating LLMs on sarcasm understanding, and reported that GPT-4 consistently outperformed other tested LLMs across prompting methods, with an average improvement of 14.0%, while all tested LLMs still underperformed supervised PLM-based baselines. In real, contextual understanding of sarcasm requires world knowledge that even large models struggle to apply consistently. Liu et al. [66] proposed CAF-I (Collaborative Agent Framework for Irony Detection), a multi-agent LLM system where specialized agents handle different aspects of irony analysis, literal meaning extraction, context modeling, and incongruity detection. CAF-I achieves state-of-the-art zero-shot irony detection with an average accuracy of 76.89%, suggesting that multi-agent decomposition can mitigate

some limitations of single-model irony detection systems. Oprea and Bâra [67] explored an LLM-as-a-judge paradigm for sarcasm detection, using LLMs to evaluate and calibrate the outputs of smaller models; their approach combining DistilBERT-SST2 with LLM-based quality scoring achieved a macro F1 of 0.8784, outperforming both standalone LLMs and standalone fine-tuned models. These results collectively suggest that sarcasm detection benefits most from hybrid architectures that combine the linguistic pattern recognition of fine-tuned models with the pragmatic reasoning capabilities of LLMs.

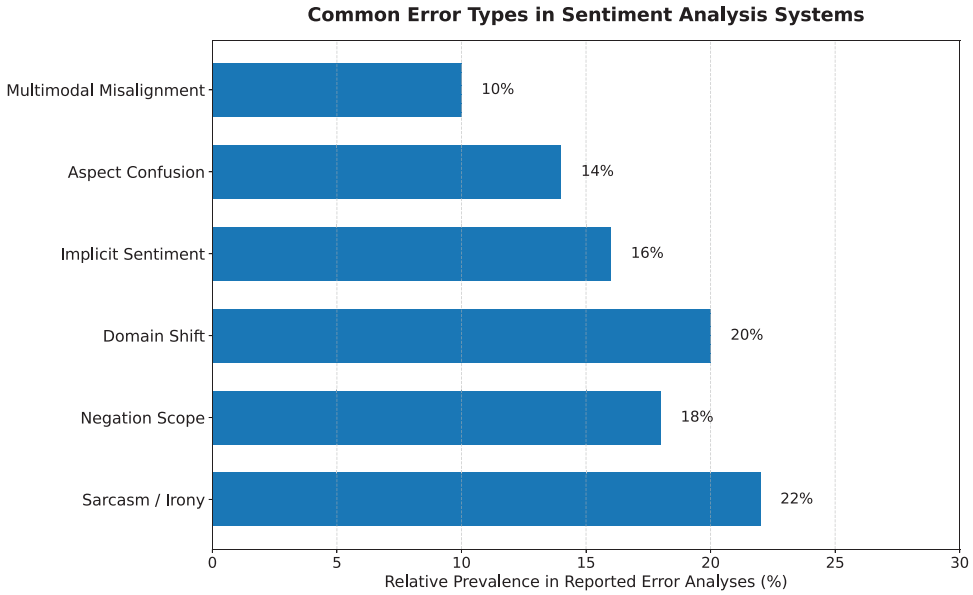

Figure 13: Common error types reported in sentiment analysis systems across benchmarks and application domains. Percentages indicate relative prevalence in published error analyses and qualitative studies, reflecting recurring challenges rather than exact empirical frequencies (This figure was generated with the assistance of AI-based tools).

5.2 Domain and Temporal Drift

Sentiment analysis models trained on one domain often perform poorly when applied to another, a phenomenon known as *domain drift* [29,30,126]. Blitzer et al. [29] documented performance drops significantly (often 10–15 percentage points) when transferring between books, DVDs, electronics, and kitchen appliance reviews, while Glorot et al. [30] proposed deep learning-based domain adaptation using stacked denoising autoencoders.

Temporal drift presents a complementary challenge: language evolves over time, and models trained on historical data may fail on contemporary text. Lazaridou et al. [127] showed that language models experience measurable degradation when evaluated on text from time periods outside their training distribution, a finding particularly relevant for social media SA, where slang, memes, and cultural references change rapidly. Sequential adaptation to new data can further exacerbate this issue through catastrophic forgetting, where previously learned knowledge is overwritten by new information [128]. The pre-trained transformer paradigm partially addresses domain drift through transfer learning, as BERT's general linguistic knowledge transfers across domains with minimal fine-tuning [19]. However, LLMs introduce a new form of temporal drift: their training data has a cutoff date, and sentiment about evolving entities (products, public figures, policies) may change post-training.

5.3 Long-Form Context Processing

Many real-world SA tasks involve long documents, including full movie reviews, earnings call transcripts, and legislative texts, that exceed the context windows of standard models [73,129,130]. BERT's 512-token limit is particularly problematic; while a typical IMDb review contains 230 words on average, the distribution has a long tail extending beyond 2000 words. Beltagy et al. [129] introduced the Longformer with a sparse attention pattern that scales linearly with sequence length, enabling processing of documents up to 4096 tokens. ModernBERT [58] further extends this to 8192 tokens with rotary positional embeddings. LLMs offer the most dramatic improvement, for example Llama 3 supports up to 128K tokens [21]. Recent long-context LLMs such as Gemini 1.5 support context windows approaching one million tokens in certain configurations. While this significantly expands the amount of text that can be processed in a single pass, practical limitations including latency, computational cost, and attention dilution which mean that context length remains a relevant consideration for large-scale sentiment analysis tasks.

5.4 Cultural and Linguistic Diversity

The vast majority of SA research has focused on English, creating a significant resource gap for other languages [31,32,86,88,131]. Cross-lingual SA aims to transfer sentiment knowledge from high-resource to low-resource languages, with XLM-R [32] serving as a foundational contribution: Conneau et al. trained a multilingual transformer on 100 languages, enabling zero-shot cross-lingual SA by fine-tuning on English data and evaluating on target languages.

Recent advances have expanded this frontier considerably. Miah et al. [61] proposed a multimodal approach to cross-lingual SA across Arabic, Chinese, French, and Italian, using transformer ensembles that combine textual and acoustic features. Zhu et al. [62] introduced a Model Arena framework that systematically compares XLM-R, Llama-3, and GPT-4 for cross-lingual SA, finding that model selection depends critically on the target language family and available supervision. Chen et al. [63] addressed the intersection of cross-lingual and multimodal SA for low-resource languages, introducing the LFD-RT framework that leverages language-family-driven resource transfer. Chen et al. [64] proposed an adaptive self-alignment method for bridging resource gaps in cross-lingual SA, demonstrating improvements for under-resourced language pairs. Šmíd and Král [33] provide one of the most comprehensive survey to date on cross-lingual aspect-based sentiment analysis, cataloging methods, datasets, and open challenges for this intersection of cross-lingual NLP and ABSA. The LACA framework [65] demonstrates a novel approach using LLM-generated data augmentation for cross-lingual ABSA, where LLMs generate synthetic training data in target languages, partially bridging the annotation gap. Finally, Horsa and Tune [78] and Musa et al. [79] represent important contributions toward SA in African languages (Afaan Oromoo and Hausa, respectively), demonstrating that transformer-based approaches can be adapted for extremely low-resource settings. As shown in Fig. 14, cross-lingual transfer remains highly effective for high-resource languages, while performance degrades substantially for low-resource settings, motivating recent African language–focused benchmarks and datasets.

Beyond linguistic resource gaps, sentiment analysis faces deeper **cultural nuance challenges** that are not resolved by cross-lingual model transfer alone. Cross-cultural communication patterns can influence how sentiment is expressed linguistically. Some studies suggest that indirect expressions of criticism may appear more frequently in certain cultural contexts, creating additional challenges for sentiment classification across languages and regions. Politeness conventions, humor norms, and the pragmatics of irony vary substantially across language communities, meaning that a model achieving high accuracy on English sarcasm may fail entirely on Japanese *tatemae* (public facade) or Arabic rhetorical understatement. Morphologically rich and agglutinative languages (e.g., Turkish, Finnish, Hausa) present compositional challenges where sentiment

modifiers attach as affixes rather than separate words, defeating word-level lexicon approaches and requiring morphology-aware architectures. Code-switching, the mixing of two or more languages within a single utterance, common in multilingual communities creates additional complexity for both tokenization and sentiment inference. These non-linguistic dimensions of cultural diversity remain largely unaddressed in the current benchmarking ecosystem and represent a critical gap for globally deployable SA systems.

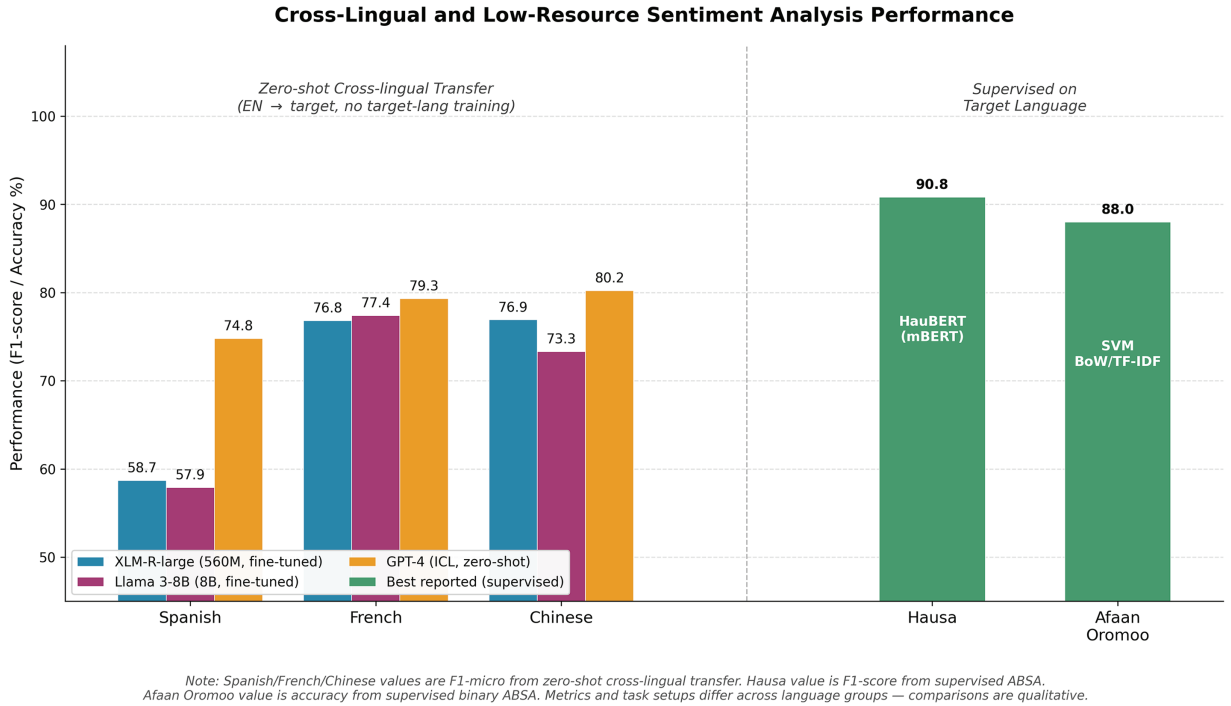

Figure 14: Cross-lingual sentiment analysis performance across language resource levels. Spanish, French, and Chinese values are F1-micro scores from zero-shot cross-lingual transfer (English source) reported in [62], using XLM-R-large (560M, fine-tuned), Llama 3-8B (fine-tuned), and GPT-4 (in-context learning, zero-shot). The Hausa value is the F1-score of HauBERT (mBERT) on supervised ABSA [79], and the Afaan Oromoo value is the accuracy of an SVM with BoW/TF-IDF features on supervised binary ABSA [78]. Illustrative values reflect trends reported across multilingual evaluations, highlighting the ~30-point degradation from high-resource to low-resource languages and the narrowing gap between fine-tuned and zero-shot approaches in high-resource settings. Metrics and task setups differ across language groups; comparisons are qualitative.

5.5 AI-Generated Fake Reviews

The proliferation of LLMs has created a new and urgent challenge: the generation of sophisticated fake reviews that are increasingly difficult to distinguish from human-written content [37,38]. This represents an adversarial dimension of SA where the same technology used for analysis is weaponized for deception. A multi-domain analysis of LLM-generated fake reviews [37] found that AI-generated open-ended reviews cause a 25%+ accuracy drop in detection systems compared to template-based fakes, because the increased fluency and coherence of LLM-generated text renders traditional statistical features, n-gram distributions, perplexity scores, far less effective. Research on ChatGPT-paraphrased reviews [38], analyzing patterns across 20 hotels with 6000 fake reviews, revealed that LLM-paraphrased reviews exhibit distinctive patterns in semantic similarity distributions that can be exploited for detection, though these patterns evolve as models improve. The resulting arms race between generation and detection defines this emerging subfield: fine-tuned BERT and RoBERTa classifiers achieve reasonable detection accuracy on current LLM outputs,

but generalization to future model versions remains an open problem. Recent regulatory developments reflect growing concern about automated content manipulation. For example, the U.S. Federal Trade Commission's 2024 rule targets deceptive or fabricated consumer reviews and testimonials, including those generated using AI when they misrepresent genuine consumer experiences.

For practitioners deploying SA in review-sensitive domains (e-commerce, hospitality, healthcare), a multi-layer detection pipeline offers the most robust current approach. The first layer employs *statistical features*: perplexity scores from calibrated language models, burstiness metrics (variance in sentence-level surprise), and distributional signatures that differ between human and machine text. The second layer uses *neural classifiers* (fine-tuned RoBERTa or ModernBERT) trained on labeled human/AI review corpora, which capture subtler stylistic differences. The third layer applies *cross-modal verification*: comparing textual sentiment against behavioral signals (purchase history, reviewer profile patterns, temporal posting patterns) to identify reviews that are linguistically plausible but behaviorally anomalous.

A critical production consideration is *detector generalization*: classifiers trained on GPT-3.5-generated text may fail on GPT-4o-generated text, necessitating continual re-training as generative models evolve. This suggests that detection systems must include online monitoring for distributional drift in review characteristics, triggering re-calibration when the detection boundary shifts. The regulatory landscape is also evolving: the U.S. FTC's 2024 rule and emerging EU AI Act provisions create compliance requirements that further motivate investment in detection infrastructure. We emphasize that provably robust defenses against adversarial text generation do not yet exist, making this an active arms race rather than a solved problem.

5.6 Bias, Fairness, and Ethics

Sentiment analysis systems can perpetuate and amplify societal biases present in training data [132,133]. Kiritchenko and Mohammad [133] evaluated 219 SA systems, finding systematic differences in sentiment scores assigned to sentences mentioning different demographic groups, while Sheng et al. [132] demonstrated that language models generate more negative sentiment text when prompted with certain demographic attributes. These biases have real-world consequences when SA systems are deployed for hiring decisions, content moderation, financial analysis, or public policy evaluation. The intersection of sentiment analysis with online content moderation represents a particularly high-stakes application, where biased sentiment classifiers can systematically over-flag content from certain demographic or linguistic communities, creating disparate impacts on speech and participation. Addressing bias requires both technical interventions (debiasing training data, adversarial training, fairness constraints) and governance frameworks encompassing audit protocols, transparency requirements, and impact assessments.

The challenges outlined above underscore that further progress in sentiment analysis is unlikely to arise from incremental architectural scaling alone. Instead, they point toward the need for fundamentally new system designs that incorporate reasoning, adaptability, privacy preservation, and human oversight. In response, a set of emerging research frontiers has begun to take shape, extending sentiment analysis beyond single-model prediction toward more structured, interactive, and trustworthy frameworks.

6 Emerging Frontiers

The emerging directions in sentiment analysis, including deployment considerations, explainability, and cross-lingual transfer, are best understood as interconnected challenges rather than isolated research threads. Across these dimensions, a fundamental trade-off persists between performance, interpretability, and stability. High-capacity models such as LLMs offer strong generalization but introduce variability and opacity, while more structured approaches provide interpretability at the cost of flexibility. Similarly, cross-lingual methods extend coverage but often amplify uncertainty due to data scarcity and cultural variation.

Understanding sentiment analysis systems through this unified lens highlights that future progress will depend not only on improving accuracy, but on achieving robust, interpretable, and deployable solutions across diverse real-world settings.

The rapid evolution of sentiment analysis, particularly in the era of large language models, has opened several research frontiers that extend well beyond incremental improvements on standard benchmarks. This section surveys the most promising of these directions.

6.1 Reasoning-Augmented Sentiment Analysis

Beyond the chain-of-thought approaches discussed in Section 4.5, the frontier of reasoning-augmented SA includes multi-step inference systems that decompose complex sentiment tasks into tractable sub-problems. The emergence of reasoning-oriented LLMs suggests a possible direction for sentiment analysis systems that perform more structured multi-step inference when mapping events, context, and expressed opinions to sentiment judgments. Multi-chain CoT for ABSA [57] represents one such advance, employing parallel reasoning chains that independently analyze different aspects before synthesizing a final judgment. Graph-enhanced approaches [117] integrate structured knowledge into the reasoning process, using graph neural networks to model relationships between aspects, opinion expressions, and context.

6.2 Agentic Sentiment Analysis Workflows

The CAF-I framework for irony detection [66] exemplifies a broader trend toward multi-agent systems for SA. In agentic workflows, specialized LLM agents handle different components of the analysis pipeline: an aspect extraction agent identifies relevant entities and aspects; a context analysis agent models situational and cultural context; a sentiment reasoning agent applies CoT reasoning to determine polarity; and a calibration agent adjusts for known biases and model variability. This decomposition allows each agent to be optimized independently and enables human-in-the-loop oversight at each stage of the pipeline.

The agentic paradigm offers particular advantages for complex SA tasks that involve multiple interacting challenges, for example, cross-lingual sarcasm detection in multimodal content, where a single model would need to simultaneously handle language transfer, pragmatic inference, and modality fusion. By decomposing this into specialized agents, each sub-task can leverage the most appropriate methodology (e.g., a fine-tuned cross-lingual encoder for language transfer, a CoT-prompted LLM for pragmatic inference, and a dedicated fusion module for multimodal integration). The orchestration of these agents, including conflict resolution when agents disagree and confidence calibration across the pipeline, represents an active research frontier with connections to the broader multi-agent systems literature.

6.3 Federated and Privacy-Preserving Sentiment Analysis

Healthcare and financial sentiment analysis (SA) applications often require processing sensitive data under strict privacy constraints. Federated learning (FL) enables training SA models across distributed datasets without centralizing raw data, by iteratively aggregating locally computed updates from multiple clients.

Formally, let K denote the number of participating clients and $\theta^{(t)}$ the global model parameters at communication round t. The global model is updated as:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{1}{K}\sum_{k=1}^{K}\Delta\theta_k^{(t)} \tag{32}$$

here, $\Delta\theta_k^{(t)}$ represents the update computed by client k after performing local optimization on its private dataset. In practice, these updates are typically obtained via one or more steps of stochastic gradient descent (SGD), i.e.,

$$\Delta\theta_k^{(t)} \approx -\eta \nabla \mathcal{L}_k(\theta^{(t)}),$$

where $\mathcal{L}_k$ denotes the local loss function and η is the learning rate. The aggregation step in Eq. (32) approximates a descent direction for the global objective $\sum_{k=1}^{K} \mathcal{L}_k$ without requiring data centralization.

The formulation above assumes equal contribution from all clients. In practice, federated averaging typically uses a weighted aggregation based on local dataset sizes:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t)},$$

where n_k is the number of samples at client k, $n = \sum_{k=1}^{K} n_k$, and $\theta_k^{(t)} = \theta^{(t)} + \Delta\theta_k^{(t)}$ denotes the locally updated model. To further enhance privacy, differential privacy (DP) can be incorporated by perturbing client updates before aggregation:

$$\Delta\tilde{\theta}_k = \Delta\theta_k + \mathcal{N}(0, \sigma^2 I) \tag{33}$$

In practice, DP is implemented by first clipping updates to a fixed norm bound C,

$$\Delta\theta_k \leftarrow \frac{\Delta\theta_k}{\max(1, \|\Delta\theta_k\|_2 / C)},$$

followed by the addition of Gaussian noise. The noise term $\mathcal{N}(0, \sigma^2 I)$ is calibrated to the sensitivity of the clipped updates, where σ controls the privacy–utility trade-off: larger σ provides stronger privacy guarantees at the cost of reduced model accuracy. Intuitively, the injected noise obscures the contribution of individual data points, making it difficult to infer sensitive information from shared updates.

In practice, federated SA introduces additional challenges beyond standard FL settings. Sentiment label distributions are often highly non-i.i.d. across clients (e.g., different institutions may exhibit distinct linguistic patterns or domain-specific sentiment cues), requiring robust aggregation strategies such as FedProx. Moreover, the privacy–utility trade-off is particularly pronounced for SA, where high-dimensional transformer representations lead to large gradient magnitudes, necessitating stronger noise injection for DP guarantees and potentially degrading fine-grained sentiment distinctions.

6.4 Real-Time and Edge Deployment

Deploying SA at scale requires efficient models, and two key techniques from the LLM efficiency literature are particularly relevant. LoRA (Low-Rank Adaptation) [134] freezes the pre-trained weights and injects trainable low-rank decomposition matrices:

$$\mathbf{W}' = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{BA} \tag{34}$$

here, $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$ denotes the frozen pre-trained weight matrix, and the low-rank update $\Delta\mathbf{W} = \mathbf{BA}$ has rank at most r, ensuring that $\mathbf{W}'$ has the same dimensionality as $\mathbf{W}_0$. $\mathbf{B} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$, reducing trainable parameters by over two orders of magnitude while maintaining competitive performance. QLoRA [135] extends extends this with 4-bit quantization, enabling memory-efficient fine-tuning of very

large models and making local or resource-constrained deployment substantially more feasible than full-precision alternatives.

Deployment considerations vary substantially with hardware, quantization level, batching strategy, sequence length, and serving framework. In general, lexicon-based systems are the most lightweight and latency-efficient; fine-tuned encoder models such as BERT-base are typically practical for high-throughput production inference; quantized mid-scale LLMs can support local or on-premise deployment with substantially higher latency and memory demands; and frontier-scale LLMs usually require either hosted APIs or multi-accelerator infrastructure. Accordingly, deployment comparisons should be reported together with the underlying hardware and inference configuration rather than as hardware-independent universal numbers.

6.5 Explainable Sentiment Analysis

As SA systems are deployed in high-stakes settings, explainability becomes critical. Danilevsky et al. [136] survey XAI for NLP, identifying attention visualization, feature attribution, and natural language explanations as primary explanation modalities. Two widely adopted frameworks deserve particular attention.

LIME (Local Interpretable Model-agnostic Explanations) [137] explains individual predictions by fitting a local linear model:

$$\xi(\mathbf{x}) = \arg\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g) \tag{35}$$

where f is the black-box model, g is an interpretable surrogate, $\pi_{\mathbf{x}}$ defines a locality kernel, and $\Omega(g)$ penalizes complexity. $\mathcal{L}(f, g, \pi_{\mathbf{x}})$ measures the fidelity of the surrogate model g in approximating the black-box model f in the locality defined by $\pi_{\mathbf{x}}$, and $\mathcal{G}$ denotes the class of interpretable models. SHAP (SHapley Additive exPlanations) provides theoretically grounded feature attributions based on cooperative game theory, computing the SHAP value for feature j as:

$$\phi_j = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} [f(S \cup \{j\}) - f(S)] \tag{36}$$

where $\mathcal{F}$ is the full feature set and S ranges over all subsets excluding feature j. $f(S)$ denotes the model output when only the features in subset S are present (with other features marginalized or fixed to baseline values), and $\mathcal{F}$ is the full set of input features.

Recent applications of XAI to SA have been diverse and promising. ModernBERT has been combined with SHAP and LIME for transparent sentiment classification [68], suggesting that modern encoder architectures can support both strong predictive performance and post hoc interpretability. Explainable ABSA using transformers with LIME, SHAP, attention visualization, integrated gradients, and Grad-CAM [69] provides a comprehensive toolkit to understand aspect-level predictions. In a domain-specific application, ABSA for software requirements elicitation with XAI [70] achieved F1 = 0.94 on the ACP dataset by combining BERT with LIME explanations, enabling software engineers to understand and trust aspect-level sentiment predictions. Rajani et al. [138] reported that generating natural language explanations alongside predictions can improve both model performance and user trust. This is increasingly relevant for LLM-based sentiment analysis, although fluency of explanation should not be conflated with explanation faithfulness or causal correctness.

To complement the largely conceptual discussion of explainability and reasoning mechanisms, Table 5 summarizes representative quantitative findings from prior studies, highlighting the measurable impact of prompt design, reasoning strategies, and model choice on sentiment analysis performance.

Table 5: Reported performance impact of explainability and prompt variability in sentiment analysis models. Values are representative ranges reported in prior studies with heterogeneous experimental conditions (different datasets, baselines, and LLM versions); direct quantitative comparison across rows is not intended.

Factor	Observed Impact	Reference Insight
Prompt variation (LLMs)	10%–15% accuracy variation	SentiEval benchmark [22]
LLM vs. Lexicon correlation	r = 0.59–0.77 vs. 0.20–0.30	Rathje et al. [24]
CoT reasoning (implicit SA)	+5%–12% improvement	THOR [54]/SAoT [55] studies
CoT on simple tasks	−2%–5% degradation	Zheng et al. [56]

We emphasize that the ranges summarized in Table 5 are drawn from studies with heterogeneous experimental setups i.e., different datasets, baselines, prompt formats, and LLM versions and the source papers do not consistently report normalized variance, confidence intervals, or significance tests for the quoted deltas. Accordingly, the reported magnitudes should be read as indicative order-of-effect estimates rather than as directly comparable measurements, and side-by-side numerical comparison between rows is not intended.

As shown in Table 5, explainability-related factors such as prompt formulation and reasoning strategies are associated with non-trivial variability. These results reinforce that explainability is not merely an interpretive layer, but an active component influencing model behavior, particularly in LLM-based systems.

We further caution that the associations in Table 5 are largely correlational: most source studies vary prompts or reasoning strategies without holding model version, decoding temperature, in-context examples, and dataset composition strictly constant, so the observed deltas conflate prompt effects with model and data effects. Controlled ablations that perturb only the prompt or reasoning scaffold while fixing the underlying model and decoding configuration remain scarce, and are a prerequisite for treating prompt engineering or chain-of-thought as reliable, isolable performance levers rather than joint artifacts of a particular deployment configuration.

6.6 Cross-Lingual Transfer and Low-Resource Languages

The cross-lingual frontier extends beyond simple model transfer to address deeper structural challenges. Šmíd and Král's comprehensive survey [33] identifies key open problems including aspect term extraction across languages with different morphological structures, sentiment compositionality in agglutinative languages, and evaluation standardization across language pairs. The LACA framework [65] represents a novel LLM-augmented approach where large language models generate synthetic aspect-based sentiment training data in target languages, effectively using the LLM's multilingual knowledge as a data augmentation engine. Chen et al. [63] further advance this direction with language-family-driven resource transfer, exploiting linguistic typological similarities between related languages. Chen et al. [64] introduce adaptive self-alignment for bridging resource gaps, where models learn to align sentiment representations across languages without parallel corpora, using self-supervised objectives on monolingual data.

Critical Assessment of Synthetic Data Approaches: While frameworks such as LACA offer pragmatic solutions for data-scarce settings, they also introduce important risks. When downstream models are trained heavily on LLM-generated sentiment data, several failure modes may arise, including reduced output diversity, amplification of biases present in the generating model, and erosion of subtle human-annotated distinctions. These concerns do not invalidate synthetic data augmentation, but they make rigorous auditing against human-annotated data essential. First, *model collapse*: iterative training on LLM-generated data can cause progressive narrowing of output distributions, as each generation cycle reinforces the generating

model's biases and reduces diversity. Second, *bias amplification*: systematic biases in the generating LLM such as sentiment skew toward certain topics, cultural biases inherited from English-centric pre-training, and stylistic homogeneity are inherited and potentially magnified by the downstream model. Third, *erosion of ground truth*: subtle human-annotated nuances, ambivalent sentiment, culture-specific pragmatic markers, implicit opinions expressed through discourse-level patterns may be systematically smoothed out by LLM-generated data that favors prototypical, unambiguous examples. These risks do not invalidate synthetic data approaches, but they demand rigorous quality auditing: distributional comparison between synthetic and human-annotated data, human validation of representative synthetic samples, and evaluation on held-out human-annotated test sets to ensure that synthetic training does not degrade performance on genuinely ambiguous or culturally nuanced inputs.

Taken together, the empirical patterns summarized in Table 5 and the methodological limitations discussed throughout Section 6 map directly onto the research agenda developed in Section 7. The 10%–15% accuracy swings across prompt formulations reported on SentiEval [22] reflect the absence of standardized, consistency-aware evaluation protocols for LLM-based SA, motivating the multi-axis evaluation framework and mandatory CIR reporting proposed in Sections 7.1 and 7.2. The asymmetric effects of chain-of-thought: +5%–12% on implicit sentiment tasks [54,55] but a 2%–5% degradation on straightforward polarity classification [56] expose the lack of controlled ablations in current LLM-SA research and underpin the call in Section 7.3 for human–AI collaborative annotation and standardized experimental reporting (exact prompts, temperature, seeds, and run counts). The large gap between LLM and lexicon correlations with human annotations (r = 0.59–0.77 vs. r = 0.20–0.30 [24]) is a performance signal, not an explanation; understanding *why* LLMs track human judgment better, and under which linguistic and cultural conditions they fail to do so, requires the causally grounded, cognitively informed SA proposed in Section 7.4. In this sense, the patterns in Table 5 are not isolated empirical curiosities but diagnostic symptoms of deeper methodological gaps: evaluation instability, uncontrolled confounding, and absent causal grounding that the next section seeks to make explicit and addressable.

7 Research Gaps and Future Directions

Despite the remarkable progress chronicled in this survey, the field of sentiment analysis stands at an inflection point where the most impactful advances are likely to come not from incremental refinements of existing methods, but from rethinking the fundamental assumptions, evaluation paradigms, and application contexts of the discipline. This section identifies key research gaps and articulates a forward-looking vision for the next generation of sentiment analysis systems. To provide actionable guidance, we organize future directions into three priority tiers: *Tier 1 (Immediate, 1–2 years)*, addressing gaps where solutions are feasible with current technology; *Tier 2 (Medium-term, 2–5 years)*, requiring significant research but with clear pathways; and *Tier 3 (Long-term, 5+ years)*, representing aspirational goals that require fundamental advances.

7.1 Integrated Evaluation Frameworks (Tier 1)

Despite the proliferation of benchmarks, the SA field lacks a unified evaluation framework that captures the full spectrum of capabilities. SentiEval [22] represents the most comprehensive effort to date, but it focuses primarily on English text and does not cover multimodal or cross-lingual settings. A truly comprehensive evaluation framework would need to jointly assess performance across granularity levels (document, sentence, aspect), robustness to domain and temporal drift, cross-lingual generalization, multimodal integration quality, computational efficiency and latency, explanation quality and faithfulness, and consistency across runs to address the MVP [25]. We envision such a framework as a "living benchmark"

that evolves with the field, continuously incorporating new domains, languages, and modalities as they become relevant. Such a framework would need to move beyond static leaderboards toward dynamic evaluations that test models under distribution shift, adversarial conditions, and low-resource constraints simultaneously, thereby providing a holistic view of model readiness for real-world deployment.

Proposed Evaluation Protocol: As a concrete first step, we recommend a multi-axis evaluation protocol that the community could adopt: (1) mandatory cross-domain evaluation, with training on one domain and testing on at least two held-out domains; (2) temporal robustness testing, with evaluation on data from time periods outside the training distribution; (3) demographic fairness auditing using the methodology of Kiritchenko and Mohammad [133], reporting sentiment score disparities across demographic mentions; (4) consistency testing across at least 5 inference runs for any stochastic model, reporting CIR alongside accuracy; and (5) multilingual evaluation covering at least one high-resource, one medium-resource, and one low-resource language. This protocol would enable meaningful cross-study comparisons and expose robustness gaps hidden by standard single-dataset evaluation.

7.2 Standardization Challenges (Tier 1)

The field suffers from inconsistent experimental practices: different papers use different data splits, preprocessing pipelines, and evaluation metrics, making direct comparison difficult. The green AI movement [123] further advocates for standardized reporting of computational costs alongside accuracy metrics, recognizing that the environmental and economic costs of training and deploying ever-larger models must be weighed against their marginal performance gains. Looking ahead, we believe the community should converge on standardized evaluation protocols that include mandatory reporting of carbon footprint, inference latency, and memory requirements alongside accuracy. Such protocols would accelerate progress by making results more comparable and would promote the development of models that achieve the best trade-off between performance and resource consumption, thereby democratizing access to high-quality SA across institutions with varying computational budgets.

7.3 Human-AI Collaborative Annotation (Tier 2)

The comparison of LLMs with human annotators [53] reveals nuanced complementarities: humans excel at context-dependent judgments requiring cultural knowledge, while LLMs provide superior consistency and throughput. Future annotation pipelines will likely combine LLM pre-annotation with human validation, requiring new quality control protocols and annotation interfaces. Yin et al. [139] established benchmarking protocols for zero-shot text classification that can serve as a template for standardized SA evaluation, though adaptation to the specific characteristics of sentiment tasks (subjectivity, granularity, domain dependence) remains necessary. We anticipate that the most transformative development in this space will be the design of interactive annotation systems in which humans and AI models work in tight feedback loops, each correcting and calibrating the other. Such systems would not only produce higher-quality labeled datasets but also yield insights into the systematic differences between human and machine understanding of sentiment, informing the design of more robust and culturally sensitive models. As an immediate action item, we recommend that all SA publications report: (a) exact data splits and preprocessing steps sufficient for reproduction; (b) inference latency and memory footprint; (c) carbon footprint estimates (using tools such as CodeCarbon or ML CO_2 Impact); and (d) for LLM-based methods, prompt text, temperature setting, and number of inference runs.

7.4 Toward Causal and Cognitively Grounded Sentiment Analysis (Tier 3)

Perhaps the most consequential gap in the current literature is the absence of *causal* sentiment analysis. Existing methods, from lexicons to LLMs, are fundamentally correlational: they identify associations between textual features and sentiment labels but cannot explain *why* a text conveys a particular sentiment or predict *what would change the sentiment*. A causal approach would enable counterfactual reasoning (e.g., "If the product had arrived on time, would the review be positive?"), opening new applications in customer experience management, policy analysis, and product design. Achieving causal SA will require integration with causal inference frameworks and may benefit from cognitive science insights into how humans form and update opinions. Concretely, this means developing models that can distinguish between sentiment triggers (the causal antecedents of an expressed opinion) and sentiment indicators (the linguistic markers that happen to correlate with polarity), a distinction that current approaches largely ignore. This connects to the sarcasm detection challenge (Section 5.1): understanding sarcasm fundamentally requires causal reasoning about the speaker's communicative intent, linking the pragmatic inference techniques developed for figurative language to the broader goal of causal SA.

7.5 Dynamic Sentiment Tracking and Temporal Intelligence (Tier 2)

Current sentiment analysis operates predominantly in a static, snapshot mode: a model processes a fixed piece of text and produces a classification. The real world, however, is dynamic. Public opinion about a product, policy, or public figure shifts over time in response to events, and these shifts often follow predictable patterns of escalation, equilibrium, and decay. The next generation of SA systems should be capable of *dynamic sentiment tracking*, monitoring evolving sentiment in streaming data through online learning and drift detection mechanisms. Such systems would need to combine temporal modeling techniques with SA, detecting not only what sentiment is expressed but when and why it changes. This vision extends to *predictive* sentiment analysis, in which models forecast future sentiment trajectories based on current trends and external events, enabling proactive rather than reactive decision-making in domains such as brand management, financial markets, and public health surveillance. This direction intersects with domain drift (Section 5.2): the temporal evolution of sentiment about a specific entity involves both the genuine shift in public opinion and the linguistic drift that changes how opinions are expressed, requiring models that can disentangle these two sources of temporal variation.

7.6 Sentiment Grounding in Real-World Outcomes (Tier 3)

A related and equally important gap is *sentiment grounding*: connecting textual sentiment to tangible real-world outcomes such as stock prices, election results, product sales, and public health metrics in a principled and validated manner. While correlations between sentiment signals and market movements have been widely reported, rigorous causal validation remains rare. Future work should establish when and under what conditions aggregated sentiment signals can serve as reliable leading indicators, and should develop calibrated uncertainty estimates for sentiment-based predictions. This requires moving beyond retrospective analyses to prospective, pre-registered studies that test sentiment-based forecasts against ground-truth outcomes. The multimodal fusion techniques from Section 4.6 could benefit healthcare applications where patient sentiment expressed through both text and vocal cues needs to be grounded in clinical outcomes, representing a cross-domain linkage between methodological advances and practical deployment.

7.7 Adversarial Robustness and Trust (Tier 2)

The emergence of AI-generated fake reviews [37,38] is a harbinger of a broader challenge: defending SA systems against deliberate manipulation. As LLMs become more capable, the sophistication of adversarial

attacks on sentiment systems will increase correspondingly, potentially encompassing not only fake reviews but also coordinated influence campaigns, synthetic social media personas, and targeted manipulation of financial sentiment indicators. Building adversarially robust SA systems will require a combination of provable defense mechanisms, continuous monitoring for distributional anomalies, and cross-modal verification strategies that triangulate sentiment signals across text, behavior, and contextual metadata. The broader goal is to develop SA systems that are not merely accurate in benign settings but demonstrably *trustworthy* in adversarial ones, a property that will be essential for deployment in critical infrastructure such as financial markets, healthcare systems, and democratic institutions. This challenge connects to the bias and fairness concerns of Section 5.6: both adversarial robustness and fairness involve distribution shift between training and deployment conditions, and techniques developed for one (e.g., distributionally robust optimization) may benefit the other.

7.8 Toward a Unified Vision

The convergence of several trends, including reasoning-augmented models, explainability frameworks, federated learning, and agentic architectures, points toward a future where SA systems are not merely accurate but also reliable, interpretable, fair, and privacy-preserving. Achieving this vision requires interdisciplinary collaboration spanning NLP, human-computer interaction, ethics, cognitive science, and domain expertise. The ultimate aspiration is a new class of sentiment intelligence systems that understand opinions with the depth and nuance of a human analyst while operating with the scale, speed, and consistency that only machines can provide: systems that can be trusted precisely because they can explain their reasoning, acknowledge their uncertainty, and adapt to the evolving landscape of human expression.

8 Conclusion

This survey has traced the evolution of sentiment analysis from its origins in lexicon-based methods through classical machine learning, deep learning, pre-trained transformers, and the current era of large language models. Our analysis reveals that sentiment analysis has matured from a binary classification task on movie reviews into a multi-dimensional research area spanning multiple granularity levels, modalities, languages, and domains.

The current state of the field is characterized by a productive tension between two paradigms: fine-tuned small language models (BERT, RoBERTa, ModernBERT) that achieve high accuracy with efficient inference, and large language models (GPT-4, Llama 3) that offer remarkable zero-shot flexibility at greater computational cost. The SentiEval evaluation demonstrates that neither paradigm uniformly dominates; fine-tuned SLMs lead on standard benchmarks while LLMs excel in few-shot and cross-domain settings. The model variability problem adds urgency to the development of reliable evaluation protocols that can account for the inherent stochasticity of LLM-based approaches.

Emerging frontiers, including chain-of-thought reasoning for implicit sentiment, multimodal LLMs, agentic workflows, cross-lingual transfer, explainable SA, and AI-generated fake review detection, collectively define a research agenda that extends well beyond incremental benchmark improvements. Addressing these frontiers will require not only algorithmic innovation but also new datasets, evaluation frameworks, and interdisciplinary collaboration between NLP researchers, domain experts, ethicist, and practitioners.

Sentiment analysis remains, after more than two decades, both a problem of enduring practical importance and a persistent scientific challenge. While the field has progressed from hand-crafted lexicons to trillion-parameter language models, this evolution has not eliminated the fundamental difficulty at its core. Accurately inferring sentiment ultimately requires understanding intent, context, and nuance—what people

mean rather than merely what they say. Addressing this gap will define the next phase of sentiment analysis research and determine its reliability in real-world applications.

Acknowledgement: The authors acknowledge the use of AI model, OpenAI GPT 5.2 for generating Figs. 1–5 and 9–13 as well as for improving the language quality and presentation of the manuscript. The authors take full responsibility for the accuracy, integrity, and originality of the content.

Funding Statement: This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia under grant no. (IPP: 543-305-2025). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Author Contributions: Conceptualization, methodology, literature review, data curation, visualization, and writing—original draft preparation were carried out by Shuvodeep De, Agnivo Gosai, and Karun Thankachan. Visualization, resources, and writing—review and editing were contributed by Ramadan A. ZeinEldin and Abdulaziz T. Almaktoom. Supervision and writing—review and editing were provided by Mustafa Bayram and Ali Wagdy Mohamed. Project administration was performed by Ali Wagdy Mohamed. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: No new data were created or analyzed in this study. All data supporting the findings of this work are derived from previously published studies, which have been appropriately cited in the manuscript.

Ethics Approval: Not Applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript

ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BoW	Bag of Words
CAF-I	Collaborative Agent Framework for Irony Detection
CBOW	Continuous Bag-of-Words
CIR	Classification Instability Rate
CMU-MOSEI	CMU Multimodal Opinion Sentiment and Emotion Intensity
CMU-MOSI	CMU Multimodal Opinion Sentiment Intensity
CNN	Convolutional Neural Network
CoT	Chain-of-Thought
DP	Differential Privacy
FL	Federated Learning
FLOPs	Floating-Point Operations
FTC	Federal Trade Commission
GloVe	Global Vectors for Word Representation
GQA	Grouped Query Attention
GRU	Gated Recurrent Unit
HAN	Hierarchical Attention Network
IMDb	Internet Movie Database
LACA	LLM-Augmented Cross-lingual ABSA

LFD-RT	Language-Family-Driven Resource Transfer
LIME	Local Interpretable Model-agnostic Explanations
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MELD	Multimodal EmotionLines Dataset
ML	Machine Learning
MLM	Masked Language Modeling
MLLM	Multimodal Large Language Model
MLP	Multi-Layer Perceptron
MMBERT	Multimodal BERT
MoE	Mixture of Experts
MSA	Multimodal Sentiment Analysis
MVP	Model Variability Problem
NLP	Natural Language Processing
NSP	Next Sentence Prediction
PLM	Pre-trained Language Model
QLoRA	Quantized LoRA
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Pretraining Approach
RoPE	Rotary Positional Embeddings
SA	Sentiment Analysis
SAoT	Sentiment Analysis of Thought
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
SLM	Small Language Model
SST	Stanford Sentiment Treebank
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
THOR	Three-Hop Reasoning
XAI	Explainable Artificial Intelligence
XLM-R	Cross-lingual Language Model–RoBERTa

References

1. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing—EMNLP; 2002 Jul 6–7; Philadelphia, PA, USA. p. 79–86. doi:10.3115/1118693.1118704.
2. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: a survey. WIREs Data Min Knowl. 2018;8(4):e1253. doi:10.1002/widm.1253.
3. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends in Inf Retr. 2008;2(1–2):1–135. doi:10.1561/1500000011.
4. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev. 2022;55(7):5731–80. doi:10.1007/s10462-022-10144-1.
5. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl Based Syst. 2021;226:107134. doi:10.1016/j.knosys.2021.107134.

6. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC); 2010 May 17–23; Valletta, Malta.
7. Esuli A, Sebastiani F. SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC); 2006 May 22–28; Genoa, Italy. p. 417–22.
8. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Medium. 2014;8(1):216–25. doi:10.1609/icwsm.v8i1.14550.
9. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Machine learning: ECML-98. Berlin/Heidelberg, Germany: Springer; 1998. p. 137–42. doi:10.1007/BFb0026683.
10. McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. In: Proceedings of the AAAI-98 Workshop on Learning for Text Categorization; 1998 Jul 26–27; Madison, WI, USA.
11. Turney PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL; 2002 Jul 7–12; Philadelphia, PA, USA. p. 417–24. doi:10.3115/1073083.1073153.
12. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25–29; Doha, Qatar. p. 1746–51. doi:10.3115/v1/d14-1181.
13. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations (ICLR); 2015 May 7–9; San Diego, CA, USA.
15. Wang Y, Huang M, Zhu X, Zhao L. Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1–5; Austin, TX, USA. p. 606–15. doi:10.18653/v1/d16-1058.
16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT); 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
17. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:190711692. 2019.
18. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems (NeurIPS). Vol. 32. Red Hook, NY, USA: Curran Associates Inc.; 2019.
19. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune BERT for text classification? In: Chinese computational linguistics. Cham, Switzerland: Springer International Publishing; 2019. p. 194–206. doi:10.1007/978-3-030-32381-3_16.
20. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Advances in neural information processing systems (NeurIPS). Vol. 33. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 1877–901.
21. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. arXiv:240721783. 2024.
22. Zhang W, Deng Y, Liu B, Pan S, Bing L. Sentiment analysis in the era of large language models: a reality check. In: Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024; 2024 Jun 16–21; Mexico City, Mexico. p. 3881–906. doi:10.18653/v1/2024.findings-naacl.246.
23. Krugmann JO, Hartmann J. Sentiment analysis in the age of generative AI. Cust Needs Solut. 2024;11(1):3. doi:10.1007/s40547-024-00143-4.
24. Rathje S, Mirea DM, Sucholutsky I, Marjieh R, Robertson CE, Van Bavel JJ. GPT is an effective tool for multilingual psychological text analysis. Proc Natl Acad Sci U S A. 2024;121(34):e2308950121. doi:10.1073/pnas.2308950121.

25. Herrera-Poyatos D, Peláez-González C, Zuheros C, Herrera-Poyatos A, Tejedor V, Herrera F, et al. An overview of model uncertainty and variability in LLM-based sentiment analysis: challenges, mitigation strategies, and the role of explainability. Front Artif Intell. 2025;8:1609097. doi:10.3389/frai.2025.1609097.
26. Joshi A, Bhattacharyya P, Carman MJ. Automatic sarcasm detection: a survey. ACM Comput Surv. 2018;50(5):1–22. doi:10.1145/3124420.
27. Ghosh A, Li G, Veale T, Rosso P, Shutova E, Barnden J, et al. SemEval-2015 task 11:sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); 2015 Jun 4–5; Denver, CO, USA. p. 470–8. doi:10.18653/v1/s15-2080.
28. Zhang Y, Zou C, Lian Z, Tiwari P, Qin J. SarcasmBench: towards evaluating large language models on sarcasm understanding. IEEE Trans Affective Comput. 2025;16(4):2560–78. doi:10.1109/taffc.2025.3604806.
29. Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL); 2007 Jun 24–29; Prague, Czech Republic. p. 440–7.
30. Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the International Conference on Machine Learning (ICML); 2011 Jun 28–Jul 2; Bellevue, WA, USA.
31. Banea C, Mihalcea R, Wiebe J, Hassan S. Multilingual subjectivity analysis using machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing—EMNLP; 2008 Oct 25–27; Honolulu, HI, USA. p. 127–35. doi:10.3115/1613715.1613734.
32. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 8440–51. doi:10.18653/v1/2020.acl-main.747.
33. Šmíd J, Král P. Cross-lingual aspect-based sentiment analysis: a survey on tasks, approaches, and challenges. Inf Fusion. 2025;120(2010):103073. doi:10.1016/j.inffus.2025.103073.
34. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion. 2017;37(95–110):98–125. doi:10.1016/j.inffus.2017.02.003.
35. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017 Sep 7–11; Copenhagen, Denmark. p. 1103–14. doi:10.18653/v1/d17-1115.
36. Yang H, Zhao Y, Wu Y, Wang S, Zheng T, Zhang H, et al. Large language models meet text-centric multimodal sentiment analysis: a survey. Sci China Inf Sci. 2025;68(10):200101. doi:10.1007/s11432-024-4593-8.
37. Narayan A, Madhu Kumar SD, Chacko AM. Trust at risk: detecting misinformation in LLM-generated product reviews and its implications for consumer behavior and platform governance. Telematics Inform Rep. 2026;21(12):100285. doi:10.1016/j.teler.2025.100285.
38. Xylogiannopoulos KF, Xanthopoulos P, Karampelas P, Bakamitsos GA. ChatGPT paraphrased product reviews can confuse consumers and undermine their trust in genuine reviews. Can you tell the difference? Inf Process Manag. 2024;61(6):103842. doi:10.1016/j.ipm.2024.103842.
39. Jain PK, Pamula R, Srivastava G. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Comput Sci Rev. 2021;41(1):100413. doi:10.1016/j.cosrev.2021.100413.
40. Raghunathan N, Saravanakumar K. Challenges and issues in sentiment analysis: a comprehensive survey. IEEE Access. 2023;11:69626–42. doi:10.1109/access.2023.3293041.
41. Tetteh M, Thushara M. Sentiment analysis tools for movie review evaluation—a survey. In: Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS); 2023 May 17–19; Madurai, India. p. 816–23. doi:10.1109/iciccs56967.2023.10142834.
42. Islam MS, Kabir MN, Ghani NA, Zamli KZ, Zulkifli NSA, Rahman MM, et al. Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. Artif Intell Rev. 2024;57(3):62. doi:10.1007/s10462-023-10651-9.
43. Bordoloi M, Biswas SK. Sentiment analysis: a survey on design framework, applications and future Scopes. Artif Intell Rev. 2023;56(11):12505–60. doi:10.1007/s10462-023-10442-2.

44. Mao Y, Liu Q, Zhang Y. Sentiment analysis methods, applications, and challenges: a systematic literature review. J King Saud Univ Comput Inf Sci. 2024;36(4):102048. doi:10.1016/j.jksuci.2024.102048.
45. Kumar M, Khan L, Chang HT. Evolving techniques in sentiment analysis: a comprehensive review. PeerJ Comput Sci. 2025;11(3):e2592. doi:10.7717/peerj-cs.2592.
46. Alahmadi K, Alharbi S, Chen J, Wang X. Generalizing sentiment analysis: a review of progress, challenges, and emerging directions. Soc Netw Anal Min. 2025;15(1):45. doi:10.1007/s13278-025-01461-8.
47. Suryawanshi NS. Sentiment analysis with machine learning and deep learning: a survey of techniques and applications. Int J Sci Res Arch. 2024;12(2):5–15. doi:10.30574/ijsra.2024.12.2.1205.
48. Bachate M, Suchitra S. Sentiment analysis and emotion recognition in social media: a comprehensive survey. Appl Soft Comput. 2025;174(3):112958. doi:10.1016/j.asoc.2025.112958.
49. Hankar M, Mzili T, Kasri M, Beni-Hssane A. Sentiment analysis survey: datasets, techniques, applications, tools, and challenges. Knowl Inf Syst. 2025;67(10):8219–65. doi:10.1007/s10115-025-02499-y.
50. Ahmad Alomari E. Unlocking the potential: a comprehensive systematic review of ChatGPT in natural language processing tasks. Comput Model Eng Sci. 2024;141(1):43–85. doi:10.32604/cmes.2024.052256.
51. Danyal MM, Khan SS, Khan M, Ullah S, Mehmood F, Ali I. Proposing sentiment analysis model based on BERT and XLNet for movie reviews. Multimed Tools Appl. 2024;83(24):64315–39. doi:10.1007/s11042-024-18156-5.
52. Shen Y, Zhang P. Financial sentiment analysis on news and reports using large language models and FinBERT. arXiv:241001987. 2024.
53. Bojić L, Zagovora O, Zelenkauskaite A, Vuković V, Čabarkapa M, Veseljević Jerković S, et al. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. Sci Rep. 2025;15(1):11477. doi:10.1038/s41598-025-96508-3.
54. Fei H, Li B, Liu Q, Bing L, Li F, Chua TS. Reasoning implicit sentiment with chain-of-thought prompting. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, ON, Canada. p. 1171–82. doi:10.18653/v1/2023.acl-short.101.
55. Duan Z, Wang J. Implicit sentiment analysis based on chain of thought prompting. arXiv:240812157. 2024.
56. Zheng K, Zhao Q, Li L. Reassessing the role of chain-of-thought in sentiment analysis: Insights and limitations. In: Advanced intelligent computing technology and applications. Singapore: Springer Nature; 2025. p. 89–100. doi:10.1007/978-981-95-0020-8_8.
57. He Y, He Z, Gu T, Gu B, Wan Y, Li M. Multi-chain of thought prompt learning for aspect-based sentiment analysis. Appl Sci. 2025;15(22):12225. doi:10.3390/app152212225.
58. Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, et al. Longer: a modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv:241213663. 2024.
59. da Silva NB, Harrison J, Minetto R, Delgado MR, Nassu BT, Silva TH. Do multimodal LLMs see sentiment? The MLLMsent framework. arXiv:250816873. 2025.
60. Liu S, Li T. A review of multimodal sentiment analysis in online public opinion monitoring. Informatics. 2026;13(1):10. doi:10.3390/informatics13010010.
61. Miah MSU, Kabir MM, Bin Sarwar T, Safran M, Alfarhood S, Mridha MF. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Sci Rep. 2024;14(1):9603. doi:10.1038/s41598-024-60210-7.
62. Zhu X, Gardiner S, Roldán T, Rossouw D. The model arena for cross-lingual sentiment analysis: a comparative study in the era of large language models. In: Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis; 2024 Aug 15; Bangkok, Thailand. p. 141–52. doi:10.18653/v1/2024.wassa-1.12.
63. Chen L, Guan S, Huang X, Wang WJ, Xu C, Guan Z, et al. Cross-lingual multimodal sentiment analysis for low-resource languages via language family disentanglement and rethinking transfer. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; 2025 Jul 27–Aug 1; Vienna, Austria. p. 6513–22. doi:10.18653/v1/2025.findings-acl.338.
64. Chen L, Shang S, Wang Y. Bridging resource gaps in cross-lingual sentiment analysis: adaptive self-alignment with data augmentation and transfer learning. PeerJ Comput Sci. 2025;11(1):e2851. doi:10.7717/peerj-cs.2851.

65. Šmíd J, Priban P, Kral P. LACA: improving cross-lingual aspect-based sentiment analysis with LLM data augmentation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. p. 839–53. doi:10.18653/v1/2025.acl-long.41.
66. Liu Z, Zhou Z, Hu M, Chen Y, Xu Z. CAF-I: a collaborative multi-agent framework for enhanced irony detection with large language models. In: Neural information processing. Singapore: Springer Nature; 2025. p. 153–68. doi:10.1007/978-981-95-4367-0_11.
67. Oprea SV, Bâra A. LLM-as-a-judge for sarcasm detection using supervised fine-tuning of transformers. J King Saud Univ Comput Inf Sci. 2025;37(10):357. doi:10.1007/s44443-025-00379-7.
68. Prabhu O, Navada SG. ModernBERT-XAI: a synergistic approach to sentiment analysis with layer-wise learning and SHAP-LIME interpretability. Syst Sci Control Eng. 2025;13(1):2600795. doi:10.1080/21642583.2025.2600795.
69. Perikos I, Diamantopoulos A. Explainable aspect-based sentiment analysis using transformer models. Big Data Cogn Comput. 2024;8(11):141. doi:10.3390/bdcc8110141.
70. Taj S, Daudpota SM, Imran AS, Kastrati Z. Aspect-based sentiment analysis for software requirements elicitation using fine-tuned bidirectional encoder representations from transformers and explainable artificial intelligence. Eng Appl Artif Intell. 2025;151(9):110632. doi:10.1016/j.engappai.2025.110632.
71. Mienye ID, Swart TG. Ensemble large language models: a survey. Information. 2025;16(8):688. doi:10.3390/info16080688.
72. Liu B. Sentiment analysis and opinion mining. San Rafael, CA, USA: Morgan & Claypool Publishers; 2012.
73. Bhatia P, Ji Y, Eisenstein J. Better document-level sentiment analysis from RST discourse parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17–21; Lisbon, Portugal. p. 2212–8. doi:10.18653/v1/d15-1263.
74. Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing—HLT; 2005 Oct 6–8; Vancouver, BC, Canada. p. 347–54. doi:10.3115/1220575.1220619.
75. McDonald R, Hannan K, Neylon T, Wells M, Reynar J. Structured models for fine-to-coarse sentiment analysis. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL); 2007 Jun 24–29; Prague, Czech Republic. p. 432–9.
76. Thet TT, Na JC, Khoo CSG. Aspect-based sentiment analysis of movie reviews on discussion boards. J Inf Sci. 2010;36(6):823–48. doi:10.1177/0165551510388123.
77. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2004 Aug 22–25; Seattle, WA, USA. p. 168–77. doi:10.1145/1014052.1014073.
78. Horsa OG, Tune KK. Aspect-based sentiment analysis for Afaan Oromoo movie reviews using machine learning techniques. Appl Comput Intell Soft Comput. 2023;2023(1):3462691. doi:10.1155/2023/3462691.
79. Musa A, Adam FM, Ibrahim U, Zandam AY. HauBERT: a transformer model for aspect-based sentiment analysis of Hausa-language movie reviews. Eng Proc. 2025;87(1):43. doi:10.3390/engproc2025087043.
80. Scaria K, Gupta H, Goyal S, Sawant S, Mishra S, Baral C. InstructABSA: instruction learning for aspect based sentiment analysis. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 720–36. doi:10.18653/v1/2024.naacl-short.63.
81. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 4040–54. doi:10.18653/v1/2020.acl-main.372.
82. Barbieri F, Camacho-Collados J, Espinosa Anke L, Neves L. TweetEval: unified benchmark and comparative evaluation for tweet classification. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020; 2020 Nov 16–20; Online. p. 1644–50. doi:10.18653/v1/2020.findings-emnlp.148.

83. Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, Australia, p. 2236–46. doi:10.18653/v1/p18-1208.
84. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; 2013 Oct 18–21; Seattle, WA, USA. p. 1631–42. doi:10.18653/v1/d13-1170.
85. Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: Proceedings of ACL-HLT; 2011 Jun 19–24; Portland, OR, USA. p. 142–50.
86. Kralj Novak P, Smailović J, Sluban B, Mozetič I. Sentiment of emojis. PLoS One. 2015;10(12):e0144296. doi:10.1371/journal.pone.0144296.
87. Bhatia G, Nagoudi EMB, Cavusoglu H, Abdul-Mageed M. FinTral: a family of GPT-4 level multimodal financial large language models. In: Proceedings of the Findings of the Association for Computational Linguistics ACL 2024; 2024 Aug 11–16; Bangkok, Thailand. p. 13064–87. doi:10.18653/v1/2024.findings-acl.774.
88. Ranasinghe T, Zampieri M. Multilingual offensive language identification with cross-lingual embeddings. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 5838–44. doi:10.18653/v1/2020.emnlp-main.470.
89. Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics—ACL; 2004 Jul 21–26; Barcelona, Spain. p. 271–8. doi:10.3115/1218955.1218990.
90. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency LP. Multi-attention recurrent network for human communication comprehension. Proc AAAI Conf Artif Intell. 2018;32(1):5642–9. doi:10.1609/aaai.v32i1.12024.
91. Lai S, Hu X, Xu H, Ren Z, Liu Z. Multimodal sentiment analysis: a survey. Displays. 2023;80(2):102563. doi:10.1016/j.displa.2023.102563.
92. Poria S, Majumder N, Mihalcea R, Hovy E. Emotion recognition in conversation: research challenges, datasets, and recent advances. IEEE Access. 2019;7:100943–53. doi:10.1109/access.2019.2929050.
93. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahwah, NJ, USA: Lawrence Erlbaum Associates; 2001.
94. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol. 2010;29(1):24–54. doi:10.1177/0261927x09351676.
95. Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. Comput Intell. 2006;22(2):110–25. doi:10.1111/j.1467-8640.2006.00277.x.
96. Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management; 2005 Oct 31–Nov 5; Bremen, Germany. p. 625–31. doi:10.1145/1099554.1099714.
97. Manning CD, Raghavan P, Schtze H. Introduction to information retrieval. Cambridge, UK: Cambridge University Press; 2008.
98. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev. 2020;53(6):4335–85. doi:10.1007/s10462-019-09794-5.
99. Goldberg Y. A primer on neural network models for natural language processing. JAIR. 2016;57:345–420. doi:10.1613/jair.4992.
100. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:13013781. 2013.
101. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems (NeurIPS). Vol. 26. Red Hook, NY, USA: Curran Associates Inc.; 2013.
102. Pennington J, Socher R, Manning C. Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25–29; Doha, Qatar. p. 1532–43. doi:10.3115/v1/d14-1162.

103. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157–66. doi:10.1109/72.279181.
104. Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26–31; Beijing, China. p. 1556–66. doi:10.3115/v1/p15-1150.
105. Nkhata G, Gauch S, Anjum U, Zhan J. Fine-tuning BERT with bidirectional LSTM for fine-grained movie reviews sentiment analysis. arXiv:250220682. 2025.
106. Towards Data Science. Transformer architecture illustration. 2022 [cited 2026 Mar 17]. Available from: https://towardsdatascience.com/wp-content/uploads/2022/12/1B0q2ZLsUUw31eEImeVf3PQ.png.
107. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12–17; San Diego, CA, USA. p. 1480–9. doi:10.18653/v1/n16-1174.
108. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems (NeurIPS). Vol. 30. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 5998–6008.
109. The AI Summer. Query, key, value attention mechanism diagram. 2022 [cited 2026 Mar 17]. Available from: https://theaisummer.com/static/56773616d30b9dcb31aa792f2d701276/3096d/key-query-value.png.
110. Kumar S. BERT architecture diagram. 2023 [cited 2026 Mar 17]. Available from: https://sushant-kumar.com/blog/.
111. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:191001108. 2019.
112. Bello A, Ng SC, Leung MF. A BERT framework to sentiment analysis of tweets. Sensors. 2023;23(1):506. doi:10.3390/s23010506.
113. Batra H, Punn NS, Sonbhadra SK, Agarwal S. BERT-based sentiment analysis: a software engineering perspective. In: Database and expert systems applications. Cham, Switzerland: Springer; 2021. p. 138–48. doi:10.1007/978-3-030-86472-9_13.
114. Penha G, Hauff C. What does BERT know about books, movies and music? Probing BERT for Conversational Recommendation. In: Proceedings of the Fourteenth ACM Conference on Recommender Systems; 2020 Sep 22–26; Virtual. p. 388–97. doi:10.1145/3383313.3412249.
115. Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. p. 3816–30. doi:10.18653/v1/2021.acl-long.295.
116. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi EH, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Advances in neural information processing systems (NeurIPS). Vol. 35. Red Hook, NY, USA: Curran Associates Inc.; 2022. p. 24824–37.
117. Li X, Wang X, Yao C, Li Y. Graph-enhanced implicit aspect-level sentiment analysis based on multi-prompt fusion. Sci Rep. 2025;15(1):17460. doi:10.1038/s41598-025-02609-4.
118. Stilwell S, Inkpen D. Explainable prompt-based approaches for sentiment analysis of movie reviews. Proc Can Conf Artif Intell. 2024. doi:10.21428/594757db.faf9e091.
119. Gu Y, Han X, Liu Z, Huang M. PPT: pre-trained prompt tuning for few-shot learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. p. 8410–23. doi:10.18653/v1/2022.acl-long.576.
120. Cai Y, Li X, Zhang Y, Li J, Zhu F, Rao L. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Sci Rep. 2025;15(1):2126. doi:10.1038/s41598-025-85859-6.
121. Ren J. Multimodal sentiment analysis based on BERT and ResNet. arXiv:241203625. 2024.
122. Khare Y, Bagal V, Mathew M, Devi A, Priyakumar UD, Jawahar CV. MMBERT: multimodal BERT pretraining for improved medical VQA. In: Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021 Apr 13–16; Nice, France. p. 1033–6. doi:10.1109/isbi48211.2021.9434063.

123. Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. Commun ACM. 2020;63(12):54–63. doi:10.1145/3381831.
124. Kreuz RJ. The use of verbal irony: cues and constraints. In: Metaphor. London, UK: Psychology Press; 2018. p. 23–38.
125. González-Ibáñez R, Muresan S, Wacholder N. Identifying sarcasm in Twitter: a closer look. In: Proceedings of ACL-HLT; 2011 Jun 19–24; Portland, OR, USA. p. 581–6.
126. Diaz F, Mitra B, Craswell N. Query expansion with locally-trained word embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany. p. 367–77. doi:10.18653/v1/p16-1035.
127. Lazaridou A, Kuncoro A, Gribovskaya E, Agrawal D, Liska A, Terzi T, et al. Mind the gap: assessing temporal generalization in neural language models. In: Advances in neural information processing systems (NeurIPS). Vol. 34. Red Hook, NY, USA: Curran Associates Inc.; 2021. p. 29348–63.
128. McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: the sequential learning problem. In: Psychology of learning and motivation. Amsterdam, The Netherlands: Elsevier; 1989. p. 109–65. doi:10.1016/s0079-7421(08)60536-8.
129. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv:200405150. 2020.
130. He J, Wang L, Liu L, Feng J, Wu H. Long document classification from local word glimpses via recurrent attention learning. IEEE Access. 2019;7:40707–18. doi:10.1109/access.2019.2907992.
131. Barnes J, Klinger R, Schulte im Walde S. Bilingual sentiment embeddings: joint projection of sentiment across languages. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, Australia. p. 2483–93. doi:10.18653/v1/p18-1231.
132. Sheng E, Chang KW, Natarajan P, Peng N. The woman worked as a babysitter: on biases in language generation. In: Proceedings of EMNLP-IJCNLP; 2019 Nov 3–7; Hong Kong, China. p. 3407–12.
133. Kiritchenko S, Mohammad S. Examining gender and race bias in two hundred sentiment analysis systems. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics; 2018 Jun 5–6; New Orleans, LA, USA. p. 43–53. doi:10.18653/v1/s18-2005.
134. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (ICLR); 2022 Apr 25–29; Online.
135. Dettmers T, Holtzman A, Pagnoni A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. In: Proceedings of the NIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. p. 10088–115. doi:10.52202/075280-0441.
136. Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P. A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing; 2020 Dec 4–7; Suzhou, China. p. 447–59. doi:10.18653/v1/2020.aacl-main.46.
137. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 1135–44. doi:10.1145/2939672.2939778.
138. Rajani NF, McCann B, Xiong C, Socher R. Explain yourself! Leveraging language models for commonsense reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 4932–42. doi:10.18653/v1/p19-1487.
139. Yin W, Hay J, Roth D. Benchmarking zero-shot text classification: datasets, evaluation and entailment approach. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov 3–7; Hong Kong, China. p. 3914–23. doi:10.18653/v1/d19-1404.