AUTOMATIC HUMOR DETECTION: A COMPREHENSIVE SURVEY FROM THEORETICAL FOUNDATIONS TO LARGE LANGUAGE MODELS

Edward Ajayi, Prasenjit Mitra
Carnegie Mellon University Africa
Kigali, Rwanda
{eaajayi, prasenjm}@andrew.cmu.edu

ABSTRACT

Automatic humor detection, the task of computationally identifying humorous content, is increasingly critical as Large Language Models (LLMs) become integrated into human communication platforms like chatbots and virtual assistants. However, understanding humor poses significant challenges for AI due to its reliance on complex context, cultural nuances, linguistic ambiguity, and multimodal cues. Current research is fragmented across different humor types, languages, modalities, and evaluation benchmarks, particularly concerning the capabilities and limitations of modern LLMs. This survey provides a comprehensive synthesis of the automatic humor detection field, tracing its evolution from foundational psychological and linguistic theories through classical machine learning, deep learning, and the recent transformer-based LLM paradigm. We organize and analyze computational methods, feature engineering techniques, benchmark datasets (text-only, multimodal, multilingual), and evaluation metrics. We critically examine LLM adaptation strategies, including fine-tuning, parameter-efficient methods (PEFT), prompt engineering, and multi-task learning, alongside developments in multimodal and cross-lingual humor understanding. Our analysis reveals that while LLMs demonstrate improved performance in capturing surface humor patterns, significant gaps persist in deep pragmatic reasoning, cultural grounding, multimodal integration, and explainability compared to human cognition. We identify key open challenges, including data scarcity, evaluation inconsistencies, the humor-offensiveness boundary, and the need for more robust, culturally aware, and interpretable models. By consolidating the field's progress and pinpointing critical limitations, this survey aims to guide future interdisciplinary research towards developing more socially intelligent and nuanced AI systems capable of genuinely understanding human humor.

1 Introduction

Automatic humor detection is the computational task of identifying humorous content in human conversation [1]. This task requires systems to interpret context, cultural references, linguistic ambiguity, and social dynamics [2, 3]. As such, it represents one of the most complex challenges in natural language understanding for artificial intelligence. Positioned at the intersection of natural language processing, psychology, linguistics, and cultural studies, humor detection exemplifies the deeply interdisciplinary nature of computational social intelligence [4]. Humor detection deserves serious attention because humor is not only a form of amusement, but also a powerful social and psychological mechanism. It fosters bonding, reduces stress, and positively influences mental well-being [4, 5]. From a computational standpoint, however, detecting humor is exceptionally challenging. Humor frequently violates conventional linguistic expectations, depends on implicit background knowledge, and often employs irony, puns, and wordplay, which further complicate algorithmic interpretation. The field has progressed substantially over the past three decades. Early research drew on theoretical models of humor[5, 6], followed by machine learning approaches based on handcrafted linguistic features[7, 8]. More recently, deep learning[9–11] and transformer-based architectures[12, 13] have advanced the state of the art, culminating in the era of Large Language Models (LLMs). As LLMs become integral to daily human interaction through chatbots, virtual assistants, and customer service platforms used by millions worldwide, their ability to recognize and respond appropriately to humor has become increasingly critical for natural communication experiences [14]. Yet, despite these advances, humor detection remains a largely unsolved problem. Current systems often fail to recognize humorous intent [15], leading to inappropriate or culturally insensitive responses that frustrate users and reduce trust in AI. These shortcomings highlight the need for a systematic, comprehensive analysis of the field's progress and challenges, which this survey seeks to provide.

1.1 Fragmentation of Current Research:

Despite decades of progress, research on humor detection remains highly fragmented. Existing works often focus narrowly on specific humor types (e.g., irony, puns)[3], single languages[16–18], or restricted datasets[12, 19–21], making it difficult to build generalized systems. In addition, evaluations across studies are inconsistent: benchmarks vary widely in size, modality, and cultural scope, preventing systematic comparison of models.

1.2 Emergence of Large Language Models:

The arrival of LLMs has transformed natural language processing, including humor detection. Yet, despite their widespread deployment in conversational AI, rigorous evaluations of LLM humor understanding remain limited. This creates an urgent need to consolidate knowledge across traditional NLP, deep learning, and LLM-based approaches.

1.3 Need for a Comprehensive Survey:

While several reviews of humor detection exist[22, 23], they tend to be limited in scope. Some focus primarily on humor style analysis and classical ML methods[22] while others concentrate on deep learning architectures[24] or specific humor subtypes[23, 25]. None, however, provide a unified synthesis that spans linguistic and psychological foundations, NLP methods, and the recent paradigm shift introduced by LLMs and multimodal architectures. This fragmentation motivates our present survey.

Contributions

In this survey, we provide what is, to our knowledge, the most comprehensive survey of automatic humor detection to date. Specifically, we:

1. Trace the historical development of humor detection, from theoretical foundations to machine learning, deep learning, and LLMs.
2. Propose a taxonomy that organizes methods, datasets, and evaluation strategies across modalities and languages.
3. Analyze limitations of current approaches, including cultural variability, ambiguity handling, and evaluation gaps.
4. Assess the role of LLMs, highlighting both their potential and their shortcomings in humor detection.
5. Identify open challenges and future research directions to guide interdisciplinary progress in the field.

1.4 Literature Selection Criteria

To ensure comprehensive coverage, this survey is based on a systematic review of over 150 publications from 1987 to 2025. Literature was primarily sourced from Google Scholar, ACM Digital Library, arXiv, ACL Anthology, Springer Nature, and IEEE Xplore using keywords such as "humor detection," "humor recognition," "computational humor," "irony detection," "multimodal humor," "humor datasets," and "humor theory," combined with method-specific terms ("machine learning," "deep learning," "large language models").

Inclusion criteria prioritized empirical studies reporting performance metrics (e.g., Accuracy, F1-score) or offering significant methodological/theoretical insights, coverage of diverse humor types, languages, and modalities, relevance to theoretical foundations (Section 2), computational techniques in humor detection (Sections 3-6), and availability in English. Foundational theoretical texts [26, 27] provide historical context, while peer-reviewed articles and preprints offer empirical grounding. Exclusion criteria filtered out non-scholarly sources, duplicates, studies lacking quantifiable results, and narrowly scoped LLM papers that did not contribute substantively to humor-specific analysis. This survey includes impactful contemporary LLM-focused studies within Section 6, while avoiding works that fall outside the scope of computational humor. The resulting corpus balances seminal contributions [26, 28, 29] with recent advancements[], addressing the field's fragmentation by incorporating diverse publication venues.

2 Theoretical Foundations of Humor

Humor represents a complex phenomenon in human communication that has been conceptualized differently across decades of research, reflecting varying theoretical understandings of this multifaceted domain. Veatch [26] defines humor as a specific psychological state that tends to produce laughter, though this definition underscores the inherent complexity of humor, as laughter can emanate from diverse cognitive and social mechanisms. Berger [30] characterized humor as an enigma because humans universally seek it across cultures and nations[31, 32], with virtually all aspects of human experience remaining susceptible to humorous interpretation. Beyond purely theoretical understanding, researchers have conducted cross-cultural investigations across different continents to examine how humor manifests in different cultural communications [33, 34]. This research demonstrates that humor, potentially more than language itself, represents a universal phenomenon common across cultures, yet its specific manifestations and interpretations resist straightforward generalization due to the complex interplay of cultural, linguistic, and social factors. Dating back to the 4th century BCE[5], systematic theoretical investigations of humor have generated distinct explanatory frameworks for understanding this phenomenon. These theoretical approaches have evolved considerably over the millennia, and this section examines the major theoretical frameworks and analyzes their relevance to contemporary computational humor detection research.

2.1 Classical Humor Theories

2.1.1 Superiority Theory

Superiority theory conceptualizes humor as fundamentally rooted in social dominance and comparative advantage, with laughter serving as an expression of perceived superiority over others. Aristotle [35] characterized comedy as based on the imitation of individuals "worse than average," establishing a framework where humor emerges from social hierarchies in which higher-status individuals mock those perceived as inferior [36]. This theoretical perspective posits that comedy and laughter result from feelings of superiority, with the underlying assumption that the targets of humor behave in foolish or inadequate ways, creating the necessary conditions for humorous perception. While commonly attributed to Plato [37], superiority theory has been developed and refined by numerous philosophers including Aristotle, Hobbes, and Morreall, each building upon Plato's foundational conceptualization [37]. The theory emphasizes that ridicule and feelings of relative superiority constitute necessary components of humor[38], with Morreall arguing that this framework adequately accounts for self-deprecating humor by positioning individuals as superior to their former selves. Contemporary formulations summarize the theory as proposing that laughter represents an expression of feeling superior either to others or to one's previous state [6, 35, 37, 39].

2.1.2 Release/Relief Theory

In contrast to superiority theory, relief theory posits that laughter, when properly regulated and proportioned, provides psychological pleasure and relaxation [5] without necessarily emanating from feelings of superiority or individual mockery. This theoretical framework shifts focus toward understanding the mechanisms that trigger laughter and the psychological processes underlying humorous responses [35, 39]. The psychological dimensions of humor prove crucial to this theory, as it proposes that humor emerges when psychological tension is released from individuals, with jokes serving as effective stimuli for inducing this tension-release response. The theoretical consensus maintains that while moderate laughter provides beneficial effects, excessive laughter indicates ridicule or mockery and therefore presents moral concerns [5]. The fundamental principle underlying release-based theories holds that laughter "provides relief for mental, nervous and psychic energy, and this ensures homeostasis after a struggle, tension, and strain" [27]. Some theorists propose that laughter arises from the sudden transformation of an expectation into nothing, contingent upon the perceiver maintaining an appropriate psychological state for humor to occur [5]. Freud represents the most influential proponent of relief theory, with his psychoanalytic framework extensively reviewed by Attardo [40].

2.1.3 Incongruity Theory

Incongruity theory represents the most widely accepted theoretical framework for explaining humor [5, 35]. Meyer [39] succinctly characterizes this theory as proposing that people laugh at what surprises them, appears unexpected, or presents oddity in a non-threatening manner. This theoretical perspective shifts focus from loud laughter to smiling as a more nuanced form of humorous expression, with philosophers of this theory accepting smiling as a normal communicative expression between individuals [5, 38]. Incongruity is formally defined as the perception and discovery of relationships between concepts that appear fundamentally unrelated. Veatch [26] conceptualized humor through three simultaneous conditions: normality, moral violation, and the concurrent occurrence of both states, although incongruent theorists simplified this framework. Theorists recognize that while incongruity represents a necessary condition for producing laughter, it proves insufficient on its own to generate humor. Therefore, cognitive processing and interpretative mechanisms are required to distinguish genuine humor from mere nonsense, suggesting that humor emerges from the resolution of perceived incongruities rather than from their simple presence, with outcomes dependent on individual interpretation. Shurcliff [41] emphasized that surprise constitutes a key ingredient for humor, reinforcing the centrality of expectation violation in incongruity-based frameworks. Incongruity theory has generated the most extensive computational research among humor theories, as evidenced by the substantial volume of computational humor studies based on incongruity principles.

2.1.4 Benign Violation Theory

Building upon Veatch's seminal work [26], which established three necessary conditions for humor to occur, McGraw and Warren [33] developed the Benign Violation Theory as a comprehensive framework for understanding humor. This theory posits that three conditions are jointly necessary and sufficient for eliciting humor: (1) a situation must be appraised as a violation, (2) a situation must be appraised as benign, and (3) these two appraisals must occur simultaneously [33, 42]. The theory addresses limitations in previous humor theories by providing a unified explanation for what makes situations humorous rather than merely offensive or mundane. McGraw and Warren [33] conducted empirical studies examining reactions to moral violations, which typically elicit disgust but can become amusing when perceived as both violations and benign. Their findings demonstrate that potentially benign moral violations tend to elicit laughter and behavioral displays of amusement, while benign moral violations produce mixed emotions of amusement and disgust. Conversely, moral violations that are not benign (*i.e.*, malign violations) elicit strictly negative emotions. These results align with Veatch's three-level framework [26] and provide insight into the psychological mechanisms underlying humor perception. The simultaneous cognitive evaluations required by the Benign Violation Theory [33] help explain why the same stimulus can be perceived as humorous by some individuals while being offensive to others, depending on their appraisal of the violation's benign nature.

2.2 Cross-Disciplinary Humor Research

While the theories in Section 2.1 explained the approaches of philosophers and psychologists in defining the humor concept, they have also been applied across different spheres of human life. This section explores humor in organizational contexts, linguistic frameworks, and cultural perspectives, emphasizing their implications for computational humor research.

2.2.1 Humor in the Workplace

Scheel [43] highlights the significance of humor in workplace psychology, stating that humor is potentially related to all aspects of work. Their comprehensive review examines the impact of humor within work teams, organizational leadership, persuasion and negotiation, team bonding, and maintaining health and happiness. Romero and Cruthirds [44] identified several advantages of workplace humor, including stress reduction, enhanced leadership effectiveness, improved communication, fostered creativity, and strengthened organizational structure. However, humor usage within team settings has documented negative aspects [39], demonstrating the duality of humor in workplace contexts [45]. In team leadership contexts, humor provides leaders with a competitive advantage when motivating employees [46]. Self-enhancing humor styles have been found to positively relate to employee psychological well-being [47]. Leaders can strategically use humor to reduce work-related tension and pressure on subordinates, contributing to effective team functioning—a hallmark of good leadership. Teams often perceive humorous leaders more favorably, with humor serving as a key quality that influences leadership perception. During business negotiations, managers and business leaders frequently employ humorous comments, defined by Banitz [48] as remarks that are nonfactual and nonserious, with careful consideration of timing. Laughter and humor are typically deployed toward negotiation conclusions. Research indicates that humor in negotiations is associated with higher financial concessions from clients [49], suggesting its utility for improving sales outcomes and revenue generation. Regarding workplace learning, Scheel [50] explored how humor in teaching impacts learning outcomes. Humor enhances attention, memory, and learning processes, with memory retention being particularly improved when humor relates directly to the subject matter being taught. This impacts both performance and creativity. The social benefits of humor include improved student morale and enhanced trust between students and instructors [51], creating a relaxed and positive learning environment. Ryota *et al.* [52] found that teacher humor relates to student mental health, recommending that educators minimize aggressive humor in favor of affinity-based humor approaches. Workplace humor has been examined in organizational psychology, management, and communication studies. Research highlights how humor contributes to teamwork, leadership effectiveness, stress relief, and negotiation dynamics. Importantly, workplace humor reflects both positive and negative outcomes: while it can build solidarity and morale, it may also reinforce exclusion or power imbalances. For computational humor, this domain underscores the need for sentiment analysis that can distinguish between supportive versus hostile humor, and for models that account for context such as hierarchy or professional setting.

2.2.2 Humor Across Cultures

While humor is a universal phenomenon across the world's cultures and languages [30], its generalization across cultural boundaries remains challenging. The formulation and perception of humor are highly dependent on the interlocutors [53], whose shared cultural context dictates what is considered amusing. An intended humorous expression that lacks cultural sensitivity can be misunderstood or, in severe cases, perceived as an offensive moral violation. This cultural specificity has led to a significant body of research dedicated to understanding the nuances of humor across various cultures and languages [33, 34, 54].

This cultural dependency presents a major challenge for automatic humor detection systems. Reflecting this complexity, prior research in this area is extensive and explores the topic from multiple perspectives. Scholarly work has conducted direct comparisons of humor across different countries and languages [55–57], and has investigated cultural differences in humor intensity and creativity [58, 59].

Other studies have focused on variations in perception and usage patterns [34, 54, 60, 61], as well as the specific functions of humor in conversational[62] and intercultural settings [63]. Further research has sought to define the functions, intent, and measurability of cross-cultural humor [56, 64–67], while also examining its potential applications across various fields [68–70].

The breadth of these studies demonstrates the deep integration of humor into all aspects of human social life, necessitating that computational humor detection research consider diverse cultural contexts and humor interpretation frameworks.

2.3 Linguistic Theories of Humor

To understand humor from a linguistic perspective, researchers have built upon the three classical theories of humor: incongruity, superiority, and relief. Arguing that these theories are not mutually incompatible, Raskin [27] developed a script-based mechanism of humor as a neutral approach that remains compatible with established theories [71]. This approach marked a significant shift from purely psychological frameworks toward a linguistic understanding of humor mechanisms and their theoretical foundations. Raskin [72] established clear criteria for humor theory development, specifying that an effective theory must be adequate, effective, formal, constructive, computable, and explicit. Building on these principles, Raskin and Attardo [73] developed the most influential linguistic theories of humor, which have become foundational to computational humor research:

2.3.1 Script-Based Semantic Theory of Humor (SSTH)

This theory represents the first formal linguistic approach to humor analysis, developed as an application of script-based semantic theory of language [27]. The SSTH defines scripts as formal semantic entities derived from established procedures of semantic analysis, incorporating both textual content and linguistic context. These scripts form the foundation of the first formal theory of contextual semantics, making SSTH uniquely applicable to humor research due to its contextual nature. The SSTH hypothesizes that a text qualifies as a single-joke-carrying text (i.e., a joke) if and only if two conditions are satisfied [29]: (a) the text is compatible, fully or partially, with two different scripts (this should not be confused with incongruity theory, as SSTH is entirely text-based rather than psychological analysis of the statement), and (b) the two scripts are opposite in a specific semantic sense. This binary opposition creates the semantic foundation for humor generation and recognition. For a joke to occur, Raskin [29] identifies five necessary components within the text:

1. A switch from the bona-fide mode of communication to the non-bona-fide mode of joke telling
2. The text of an intended joke
3. Two (partially) overlapping scripts compatible with the text
4. An oppositeness relation between the two scripts
5. A trigger, obvious or implied, that switches from one script to the other

While the theory provides a comprehensive framework for understanding the tasks a joke maker faces when improvising humor, it remains heavily dependent on the speaker's capabilities and cannot be directly implemented as an algorithmic approach to joke generation [29] however it has been proven to be useful in various languages and culture emphasizing its importance in the field of humor study[73].

2.3.2 The General Theory of Verbal Humor (GTVH)

Recognizing that jokes represent only a limited subset of humorous texts, researchers identified significant limitations in SSTH's applicability. There are problems with applying SSTH to non-joke texts, particularly in handling larger scripts. Similarly, in an attempt to extend SSTH theory to non-joke texts, it was demonstrated that SSTH fails to provide adequate tools for analyzing features that characterize texts beyond traditional jokes, thus limiting its potential for generalization [74].

In response to these limitations, Attardo [73] developed the GTVH with the explicit intent of "creating a hub of research to describe the structure and nature of humorous texts, beginning with the simpler ones (jokes) and building up to larger, more complex ones" [74]. The GTVH extends beyond SSTH by accounting for any type of humorous text, shifting focus from purely semantic analysis to broader linguistic features.

The GTVH introduces five additional Knowledge Resources (KRs) to complement script opposition, resulting in six comprehensive KRs: Script Opposition, Logical Mechanism, Target, Narrative Strategy, Language, and Situation. This expanded framework enables the theory to handle verbal humor analysis, joke similarity assessment, and homology identification between jokes and other textual forms. However, Attardo [40] acknowledges that while GTVH addresses several limitations of SSTH, it was not specifically designed to solve the computational challenges of analyzing longer texts.

2.3.3 Ontological Semantic Theory of Humor (OSTH)

The Ontological Semantic Theory of Humor (OSTH) emerges from Ontological Semantics (OST), a comprehensive framework for natural language understanding. OST is a "theory, methodology, and, especially, technology for representing natural language meaning, for automatic transposition of text into the formatted text-meaning representation (TMR)" [75]. At its core is a language-independent ontology, which functions as a structured, engineered model of reality, meticulously defining concepts, their properties, and the intricate relationships between them [75, 76]. This central ontology is complemented by language-specific lexicons that map word senses to the concepts within the ontology.

For humor detection, OSTH applies this deep meaning representation to computationally implement the principles of script opposition from SSTH[77]. The system processes a text and translates it into one or more TMRs. Humor is detected if the system can generate at least two distinct TMRs that are compatible with the text but are in opposition to each other. The oppositeness is not merely symbolic but is formally defined within the ontology (e.g., pleasure vs. pain, expected vs. unexpected). By moving beyond statistical or keyword methods to a direct, comprehensive access to meaning, OSTH provides a robust, knowledge-based framework for computationally identifying and explaining humor in text, tackling the complexities that other theories were not designed to handle automatically [75].

3 Feature Engineering in Humor Detection

This section examines computational implementations grounded in the humor theories discussed above, presenting a systematic analysis of how researchers have operationalized theoretical frameworks into techniques for detecting humor within language, text, and human communication for automatic humor detection. These features are based on both psychological and linguistic theories of humor and have been extensively explored in humor detection research.

3.1 Theory-Based Features

Theory-based features are predominantly employed in research utilizing traditional machine learning algorithms. These approaches leverage established psychological and linguistic frameworks to create computational representations of humor.

Incongruity Detection. Incongruity theory, one of the most widely accepted frameworks of humor [78], posits that humor arises when there is a discrepancy between the listener's expectation and the actual outcome, provided that the listener can still reconcile the intended meaning [43]. This phenomenon may occur within a single sentence [79] and has been widely adopted as a feature in computational humor recognition [78].

Researchers have applied incongruity-based features across diverse domains, such as product question answering [79], product reviews [80], online conversations [81], and interpersonal studies, including humor's role in marriage [82]. Further significant contributions include computational frameworks and models leveraging incongruity for humor detection [1, 83–86]. Notably, Mihalcea et al. [83] reported that incongruity-based features outperformed alternatives, as they effectively captured surprise—an element commonly observed in humorous expression.

Ambiguity Detection. Ambiguity refers to the presence of multiple possible interpretations within a joke, allowing audiences to derive humor from unexpected or alternative meanings. Prior research has investigated ambiguity-based features through homonym detection [87], the analysis of ambiguity's role in humor recognition [88], contextualized representation approaches [89], and multimodal fusion techniques [90, 91]. Additional studies have explored the role of homographic ambiguity [92], linguistic theories of ambiguity in humor [93], and ambiguity in computational humor datasets [1].

Despite this body of work, the effectiveness of ambiguity as a discriminative signal for humor detection remains debated. Barbieri et al. [81] proposed a linguistically motivated set of features for humor detection on Twitter, highlighting ambiguity as an important factor. In contrast, Reyes et al. [94], in their study of one-liner humorous data, reported that ambiguity did not play a significant role in humorous expression. These contrasting findings suggest that the contribution of ambiguity to humor detection may be context-dependent, and further research is required to clarify its impact.

Emotion-Based Detection. Emotional cues are often leveraged in humor detection, as certain words, character combinations, and textual symbols can represent facial expressions, feelings, moods, and attitudes [22]. Prior work has integrated emotions into computational humor models through corpus-based approaches [95], emotional feature analysis [94], homographic representation studies [96], and multimodal emotion recognition frameworks [97].

However, researchers have emphasized that emotion detection from text alone is insufficient. Jain et al. [98] and Acheampong et al. [99] highlighted the need to incorporate contextual or multimodal information for more robust performance. This argument is further supported by Bijoy et al. [100], who demonstrated that multimodal fusion consistently improves emotion recognition. Similarly, Bedi et al. [101] observed that multimodal signals—such as facial expressions, prosodic cues, and speech patterns—often provide auxiliary yet crucial evidence for detecting sarcasm and humor. In some cases, these non-textual cues serve as the only reliable indicators of humorous or sarcastic intent.

Subjectivity and Contextual Information. Humor perception is inherently subjective, as what one individual finds humorous may not be perceived the same way by another [102]. This variability is particularly evident in jokes that draw on cultural beliefs, social criticism, or personal opinions. Consequently, subjectivity and contextual information play a central role in computational humor recognition. Key dimensions of contextual information include cultural background, wordplay and linguistic variation, shared knowledge, social interactions, stereotypes, and incongruity-based cues [22].

Research in this area has sought to capture these dimensions in various ways. Zhang et al. [103] investigated the role of cultural and social context in shaping humor perception, while Wiebe et al. [104] examined the integration of subjectivity features, such as opinion words and sentiment polarity, into text classification frameworks. Liu et al. [105] further advanced this line of work by modeling contextual embeddings that account for cultural and situational nuances in humorous text. Together, these studies suggest that subjectivity and context are indispensable for systems aiming to approximate human-like humor recognition. However, challenges remain in formalizing these highly variable features, particularly in cross-cultural and multilingual settings where shared background knowledge cannot be assumed.

Other Theoretical Features. Beyond ambiguity, incongruity, emotion, and subjectivity, several additional theoretical features have been explored in humor detection. One such feature is *negation*, typically expressed through words such as "not," "isn't," or "don't." Negation alters semantic polarity and can generate humorous effects when expectations are reversed or contradicted [87, 97]. Another relevant feature is *unexpectedness*, which occurs when humor emerges from absurd or implausible scenarios that catch the listener off guard, yet remain comprehensible [81, 94].

These features illustrate the richness of linguistic and cognitive signals that can contribute to humor detection. However, unlike incongruity or ambiguity, they are less frequently modeled in isolation, often appearing as part of broader feature sets. This suggests that while negation and unexpectedness capture important aspects of humorous expression, their role may be more complementary than foundational. Future work could benefit from deeper investigations into how these features interact with core humor theories, particularly in large-scale neural models where such signals are often implicitly encoded.

3.2 Lexical Features

Humor in texts manifests through a variety of lexical and structural properties, many of which are grounded in linguistic theories of humor (Section 2.3). Research in this area spans both traditional and social-media-based texts [78, 106].

Phonetic Features. Phonetic features exploit sound patterns to produce incongruity and comic effects. Examples include alliteration, where adjacent words share the same initial sounds, creating an unexpected rhythm [78, 107], and rhyme, in which words ending with the same syllables occur together [107]. Other phonetic devices, such as puns based on homophones and playful alterations of pronunciation, leverage auditory perception to enhance humor [1, 22, 87, 108].

Lexico-Semantic Features. These features focus on meaning and semantic relationships. Semantic ambiguity, including puns and double entendres, often creates humorous incongruities [106]. Emojis and symbols convey affective content, enhancing humorous interpretations [106, 109–111]. Domain-specific humor themes can be identified using lexical resources such as WordNet Domains [76, 87]. Additionally, lexical chains, which connect words via semantic relations (synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy), can capture subtle semantic patterns indicative of humor [112].

Morpho-Syntactic Features. Humorous texts often rely on unexpected structures, which can be captured through morpho-syntactic analysis. Features include part-of-speech (POS) tag ratios (e.g., nouns, pronouns, verbs, and modifiers) and POS pattern chains reflecting recurring structural sequences [106, 113]. These features are particularly effective for short texts, such as tweets or one-liner jokes, where sentence-level parsing may be unreliable.

Pragmatic and Orthographic Features. Orthographic cues, such as capitalization, exaggerated punctuation (e.g., "!!!"), and elongated words (e.g., "soooo funny"), together with pragmatic markers like discourse-level emphasis, contribute to perceived humor [78]. These features often interact with lexical, phonetic, and morpho-syntactic patterns, producing multi-level incongruities characteristic of humorous texts.

Overall, lexical features provide a set of computationally tractable cues that capture sound, meaning, structure, and stylistic patterns. These features are widely used in humor recognition systems and serve as the foundation for both traditional and modern approaches to automated humor detection.

3.3 Automated Features

Text processing through various forms of representation, including vector-based encodings and embeddings, has become crucial in humor recognition research. These approaches enable numerical encoding of textual data, capturing underlying patterns and semantic relationships between words, phrases, and sentences. Key types of automated textual features include:

Sparse / Non-Semantic Vectors. Sparse vector representations encode text as high-dimensional vectors that capture lexical information without representing semantic similarity. Examples include Bag-of-Words (BoW), which counts word occurrences without considering grammar or context [114], and TF-IDF vectors which represents text as sparse vectors where each term is weighted by its frequency in a document and inversely by its frequency across the corpus [115]. BoW and TF-IDF have been used in humor detection for various applications, including satire detection [28], social media humor [116], and Yelp review analysis [117]. Each dimension corresponds to a unique word or n-gram in the vocabulary, and most entries are zero for any given document, resulting in a sparse representation. These features are useful for detecting lexical patterns, word repetition, or stylistic markers in humor, but they do not capture contextual meaning or semantic relationships. Consequently, while historically important, sparse vectors are no longer the most commonly used approach for modern humor detection [118].

Word / Contextual Embeddings (Semantic). Dense word embeddings represent words as continuous vectors in a lower-dimensional space, capturing both semantic and syntactic relationships. Classical embeddings, such as Word2Vec and GloVe, have been applied in humor detection, including multilingual humor and irony detection [119, 120]. Modern contextualized embeddings, such as BERT, produce representations that vary depending on surrounding context and have been used in a variety of humor detection tasks [118, 119]. Sentence- or document-level embeddings, including Sentence-BERT and the Universal Sentence Encoder, capture semantic meaning at higher levels of granularity and have been applied to wordplay detection [121] and funniness assessment [122]. These embeddings are particularly effective for modeling nuanced humor, figurative language, and context-dependent jokes.

Multimodal Embeddings. Multimodal embeddings integrate textual information with other modalities, such as images, audio, or video, to produce richer representations [123]. They have been applied in humor detection from memes [123, 124], videos [91, 100], and interviews [90], where non-textual cues—such as facial expressions, speech prosody, or visual context—provide important signals for detecting sarcasm or humor that may not be inferred from text alone.

Overall, automated feature representations in humor detection have evolved from sparse, non-semantic vectors, which capture lexical patterns but lack contextual understanding, to dense word and contextual embeddings that encode semantic and syntactic relationships, and further to multimodal embeddings that integrate textual and non-textual signals. This progression has enabled more nuanced detection of humor, particularly in cases involving wordplay, figurative language, and cross-modal cues such as images, audio, and video.

Table 1: Overview of humor detection dataset categories with representative examples and key characteristics.

Category	Examples	Scale	Key Challenge
Short Text	One-liners [4], Puns [96], Tweets [106]	2K–50K samples	Context stripping, domain artifacts
Long Text	Reviews [79], Satire [28]	10K–30K samples	Discourse tracking, world knowledge
Multimodal	Memes [129], Videos [100], Stand-up [19]	5K–50K samples	Cross-modal alignment, annotation complexity
Monolingual	Spanish HAHA [59], Chinese Chumor [16]	20K–30K samples	Cultural context, translation limits

4 Datasets and Benchmarks for Humor Detection

Datasets play an outsized role in computational humor research because what constitutes humor is heavily shaped by cultural, linguistic, and multimodal context. Unlike sentiment analysis or topic classification, humor is neither universal nor easily agreed upon: annotators frequently disagree on what is funny [125], jokes may fail outside their cultural setting [53], and multimodal cues such as timing, intonation, and visual surprise are difficult to capture in text alone. As a result, benchmark design has become one of the primary bottlenecks in advancing humor detection systems.

Recent systematic reviews have catalogued humor detection datasets comprehensively [22, 78]. Building on this foundation, we provide updated coverage emphasizing theoretical alignment with Section 2, multimodal resources, and emerging LLM evaluation benchmarks. Our analysis synthesizes dataset characteristics across multiple dimensions and provides critical assessment of how dataset design choices impact system development and evaluation.

4.1 Overview of Dataset Categories

Existing humor detection resources span diverse modalities, languages, and humor types. We organize datasets along four primary dimensions that reflect both historical development and contemporary research priorities:

Modality. Datasets range from text-only resources (tweets, jokes, reviews) to multimodal corpora incorporating images (memes), audio (stand-up comedy), and video (comedy clips, interviews). This progression reflects the recognition that humor often emerges from cross-modal interactions rather than linguistic content alone.

Linguistic scope. While English remains dominant, recent efforts have produced substantial resources in Spanish [17], Chinese [16], and other languages, alongside truly multilingual benchmarks [19]. These resources are essential for studying how humor mechanisms vary across linguistic and cultural contexts.

Humor type. Datasets target specific phenomena including puns [96], satire [28], sarcasm[95, 101], irony[81, 119], and general conversational humor[101, 126]. This specialization reflects different theoretical mechanisms (ambiguity vs. incongruity vs. contextual knowledge) discussed in Section 2.1.

Task formulation. Benchmarks support binary classification (humorous vs. non-humorous), multi-class humor type identification[81, 127], and funniness scoring[128]. These formulations test different aspects of humor understanding and generation capabilities.

Table 1 provides a high-level summary of major dataset categories, which we examine in detail in subsequent subsections.

4.2 Text-Only Datasets

Text-based humor datasets form the foundation of computational humor research, with development spanning nearly two decades. These resources operationalize linguistic theories from Section 2.3 and provide controlled environments for isolating specific humor mechanisms.

4.2.1 Short-Form Text Datasets

Short-form datasets emphasize linguistic devices and enable rapid experimentation but sacrifice contextual richness. Early work by Mihalcea and Strapparava [4] established the dominant paradigm: 16,000 one-liner jokes scraped from humor websites paired with non-humorous sentences from news sources for binary classification. This dataset operationalizes incongruity theory (Section 2.1.3) by contrasting joke structures with factual reporting, though the artificial negative construction introduces exploitable artifacts [**?**].

Pun detection datasets further isolate ambiguity-based humor mechanisms. Diao et al. [1] explored homographic puns datasets where a single word carries multiple meanings, requiring models to detect semantic ambiguity discussed in section 2.3. The Pun of the day datasets used in their work achieves 0.618 Annotation Agreement Ratio (AAR) agreement which is likely because pun identification is more objective than general humor assessment.

Twitter-based datasets [17, 106] achieve larger scale (20K–50K tweets) by leveraging hashtags like #humor for distant supervision. These introduce platform-specific characteristics including informal language, emojis, abbreviations, and multimodal references, making them valuable for studying humor in social media contexts but limiting generalization to formal text.

4.2.2 Long-Form Text Datasets

Long-form datasets capture discourse-level humor requiring sustained context tracking and world knowledge. Product review datasets [79] contain approximately *20k* product reviews annotated for humor presence, where humor often appears in the form of creative complaints, unexpected comparisons, or exaggerated descriptions. These datasets test whether models can distinguish genuine humor from hyperbole or mere sentiment.

Satirical news detection [28] represents a distinct challenge: satire mimics serious news format while conveying absurd or exaggerated content, requiring models to detect incongruity between form and content rather than explicit joke structures. Successful detection often depends on recognizing implausible scenarios or inconsistent information rather than linguistic markers, connecting to the benign violation theory (Section 2.1.4) where the violation is semantic rather than structural.

Conversational humor represents a particularly important category as it reflects how humor manifests in natural human interaction. Researchers have collected conversational humor data from various sources, notably social media platforms like Twitter [17, 106], where conversations often exhibit code-mixing between multiple languages [95, 130]. Wu et al.[126] collected a large-scale multimodal conversational dataset containing multiple speakers engaged in humorous dialogues. Additionally, language-specific datasets have been developed for Spanish[17] and Chinese [16, 58], capturing humor across different conversational contexts including family jokes and everyday interactions.

4.3 Multimodal Datasets

Humor manifests in various forms of human communication, encompassing both verbal and non-verbal categories that may be expressed through audio, video, text, or combinations of these modalities. While textual humor data can be studied in isolation, humor typically emerges from multimodal interactions among speakers [126, 131]. Consequently, recent research has shifted toward multimodal humor detection, as these datasets capture crucial humor cues across different communication channels. Several multimodal humor datasets have been developed to support this research direction. Patro et al. [132] annotated episodes from the sitcom *Big Bang Theory*, focusing on laughter cues. The MUMOR dataset [126] comprises dialogues extracted from two television sitcoms. The Passau-SFCH corpus [21] captures spontaneous humor from football coaches, while M2H2 [133] draws from a popular television series. Additional datasets have been created for humor sensing applications [134]. Recent efforts continue to expand the landscape of multimodal humor datasets, including work by Ryan et al. [135], Shuo et al. [123], Bijoy et al. [100], and Barriere et al. [19]. Annotating multimodal humor datasets presents significant challenges due to the subjective nature of humor and the complexity of capturing synchronized multimodal signals. A key challenge lies in addressing the diverse domains and types of multimodal humor. Nevertheless, the proliferation of new datasets in recent years demonstrates encouraging progress in this area.

4.4 Humor in Languages

Humor is a linguistically nuanced concept that requires sophisticated language understanding, and non-native speakers may struggle to appreciate its subtleties [136]. However, data collection efforts for computational humor research have been conducted across various languages, though these efforts have predominantly focused on high-resource languages. The resulting datasets encompass diverse humor-bearing activities and contexts. Conversational humor datasets have been developed for several languages, including Latin American Spanish from TEDx talks [137], Hindi conversations [133], Chinese discussions from Ruo Zhi Ba (a Reddit-like platform) [16], and Chinese sitcoms [138]. Social media platforms have been particularly rich sources of humorous content, yielding datasets such as Spanish tweets [97, 139, 140], code-mixed Hindi tweets [95], and Persian tweets [141]. Additionally, researchers have compiled datasets featuring Chinese memes [142], Chinese jokes [143, 144], and Hindi web series conversations [145]. The LS-FUNNY dataset [137] provides another multilingual resource for humor research. Despite these efforts, the distribution of available humor datasets remains imbalanced. English dominates the landscape, followed by Chinese and other Asian languages. While European Spanish has received some coverage, there is a notable absence of humor datasets for African languages, highlighting a significant gap in computational humor research that warrants attention from the research community.

Table 2: Comprehensive overview of humor detection datasets across languages and modalities. Each entry lists its main modality combination, humor type, dataset size, and domain.

Dataset Name	Year	Lng	Modality	Humor Type	Size	Source/Domain	Ref
Text-Only Datasets							
16K One-liners	2006	EN	Text	General jokes	16K	Web scraping	[4]
Pun of the Day	2015	EN	Text	Puns/ambiguity	2.4K	Pun websites	[1]
Twitter #humor	2014	EN	Text	General	20K	Twitter	[106]
HAHA Challenge	2019	ES	Text	General	24K	Twitter	[17]
Spanish Tweets	2018	ES	Text	General	Varied	Twitter	[139]
Code-mixed Corpus	2018	HI-EN	Text	Sarcasm	5K	Twitter	[95]
Humorous Reviews	2020	EN	Text	Product reviews	19K	Amazon	[79]
Satirical News	2014	EN	Text	Satire	2K	News sites	[28]
Chumor 2.0	2024	ZH	Text	General	3K	Online forums	[16]
Chinese Jokes	2019	ZH	Text	Jokes	Varied	Web	[143]
Chinese Memes	2022	ZH	Text	Meme captions	Varied	Social media	[142]
Russian Jokes	2025	RU	Text	Spontaneous	330k	Jokes	[146]
Persian Tweets	—	FA	Text	General	Varied	Twitter	[141]
Multimodal Datasets							
UR-FUNNY	2019	EN	T+A+V	Laughter pred.	16.4K	TED talks	[131]
MUMOR	2021	EN	T+A+V	General	29.5K utt.	Sitcom	[126]
D-HUMOR	2025	EN	T+Im	General	4.3K	Memes	[147]
Big Bang Theory	2021	EN	T+A+V	Laughter cues	Varied	TV sitcom	[132]
M2H2	2021	HI	T+A+V	Conversational	6.2K utt.	TV shows	[133]
ManzaiSet	2025	JP	V	Conversational	2.3K FR.	Comedy	[148]
MUCH	2024	ZH	T+A+V	Conversational	34.8K utt.	Chinese Sitcom	[149]
AMHUSE	2017	EN	A+V+Sensors	Amusement	36 subjects	Lab exp.	[134]
Passau-SFCH	2025	DE	T+A+V	Spontaneous	Varied	Sports	[21]
StandUp4AI	2025	Multi	T+A+V	Stand-up	330 hrs	Performances	[19]
HUMEMES	2025	ZH	T+A+V	Memes	5K	Online media	[123]
Chinese Sitcom	2025	ZH	T+A+V	Conversational	Varied	Sitcom	[138]
PixelHumor	2025	EN	T+Im+A	General	2.8K	Mixed media	[135]
MemeBlip2	2025	EN	T+Im	Memes	Varied	Social media	[129]
LS-FUNNY	2025	ES	T+A	TEDx humor	Varied	TEDx talks	[137]
HumourHindiNet	2024	HI	T+A+V	Web series	Varied	Web media	[145]
HumorDB	2024	EN	Image	Visual humor	3.5K pairs	Curated images	[150]
New Yorker Caption	2024	EN	T+Im	Caption contest	2.2M captions	Cartoon contest	[151]

Key: Lng = Language; EN = English, JP = Japanese, ES = Spanish,Russian = RU, HI = Hindi, ZH = Chinese, FA = Persian, DE = German, Multi = Multilingual; T = Text, A = Audio, V = Video, Im = Image; utt. = utterances; FR = facial recordings; exp. = experiments; pred. = prediction; Ref = Reference.

4.5 Cross-Dataset Analysis and Trends

Table 2 reveals several important patterns in humor dataset development over the past two decades. First, we observe a clear temporal shift from text-only to multimodal resources. Early datasets (2006–2015) focused exclusively on textual humor, while recent efforts (2019–2025) increasingly incorporate audio and video modalities. This progression reflects growing recognition that humor comprehension requires integrating multiple communication channels, particularly for conversational and spontaneous humor where timing, prosody, and visual cues play critical roles.

Second, dataset scale varies significantly by modality and language. Text-only English datasets achieve substantial scale through web scraping and social media harvesting. In contrast, multimodal datasets remain smaller due to annotation complexity and copyright constraints on video content. Language-specific datasets for non-English languages show similar scale patterns, with Chinese datasets (Chumor 2.0: 3K samples) achieving comparable size to English resources, while datasets for lower-resource languages remain limited in both quantity and scale.

Third, humor type specialization has increased over time. Early datasets targeted general humor recognition, while recent efforts focus on specific phenomena: sarcasm [95], spontaneous humor [21], or laughter prediction [131]. This specialization enables more targeted investigation of particular humor mechanisms but fragments the research landscape, making cross-dataset comparison challenging.

Finally, source diversity remains limited. Television sitcoms dominate multimodal datasets (MUMOR[126], Big Bang Theory[132], M2H2[133], Chinese Sitcom[138]), while Twitter dominates text-only resources[81, 118]. This concentration introduces systematic biases: sitcom humor is scripted and professionally produced, while Twitter humor is optimized for virality and brevity. Genuine conversational humor in naturalistic settings remains underrepresented despite its theoretical and practical importance.

4.6 Dataset Accessibility and Reproducibility

Dataset availability significantly impacts research progress and reproducibility. While most datasets listed in Table 2 are publicly available [12, 16, 150], their access mechanisms vary significantly. This variability creates barriers to entry and hampers reproducibility efforts. For example, some datasets require direct author contact [21, 137], while others have become unavailable over time or are difficult to access despite claims of public availability [123, 126, 147].

Copyright and privacy concerns introduce further complications, particularly for multimodal datasets [21]. Content from commercial sources, such as sitcoms (e.g., [126, 149]) or specialized performances [148], often cannot be redistributed directly. This forces researchers to provide only annotations, requiring users to obtain and process the original copyrighted content separately [137]. This approach introduces versioning issues and access barriers for researchers without institutional subscriptions. Datasets based on social media content (e.g., [17, 95, 106]) face additional privacy concerns, particularly as platforms restrict API access and users delete content over time.

The lack of standardized data formats across datasets further complicates reproducibility. Text datasets vary in preprocessing, tokenization, and encoding choices. Multimodal datasets differ in temporal alignment conventions, feature extraction methods, and annotation schemas. Establishing community standards for dataset format, documentation, and distribution would significantly benefit the field.

5 Computational Approaches to Humor Detection

5.1 Overview of Computational Paradigms to Humor Detection

For centuries, psychologists and linguists have endeavored to define and understand the concept of humor, a task that has proven exceptionally challenging [5, 26]. Psychological theories date back to the 4th century BCE, providing foundational insights into the mechanisms of laughter and amusement [5, 35]. In recent decades, linguists such as Attardo and Raskin have translated these theories into computational concepts, focusing on grammars and semantics that can be deduced from languages, primarily through text [71, 73]. Natural Language Processing (NLP), as a field dedicated to computationally handling language-related tasks, has provided means for representing these concepts through features such as n-grams, bag of words, hashtags[24], bolstered by the increasing availability of datasets [22, 78]. Over the years, in a bid to perform computational humor detection—an inherently NLP task—researchers have employed various methods. These range from traditional machine learning techniques, which relied heavily on handcrafted features, to the advent of neural networks and deep learning architectures that enable more automated feature extraction and contextual understanding [22, 24, 78]. This subsection provides an overview of these computational paradigms, tracing their historical progression and interdisciplinary influences from psychology and linguistics. We highlight how early approaches laid the groundwork for modern systems, while setting the stage for discussions on classical machine learning, neural architectures, multimodal integration, and the transformative role of transformer-based models in subsequent subsections.

5.2 Classical Machine Learning Methods

Early computational approaches to humor detection relied heavily on classical machine learning algorithms, leveraging manually engineered features derived from linguistic and humor theories (discussed in Section 3). This section details the application and performance of these foundational methods.

5.2.1 Traditional Classifiers

Algorithms such as Support Vector Machines (SVMs), Logistic Regression, Naive Bayes, and Decision Trees formed the initial toolkit for humor classification tasks. These supervised learning methods learn decision boundaries or probabilistic models from labeled data (humorous vs. non-humorous text) represented by feature vectors.

Support Vector Machines (SVM) Support Vector Machines (SVMs) are supervised learning algorithms primarily used for classification, although they can be adapted for regression [152]. The core idea is to find an optimal hyperplane in a high-dimensional feature space that best separates data points belonging to different classes (e.g., humorous vs. non-humorous). This hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each class, leading to better generalization. SVMs can handle non-linear relationships effectively through the kernel trick, which implicitly maps the data to a higher-dimensional space where linear separation might be possible using functions like polynomial, radial basis function (RBF), or sigmoid kernels. In humor detection, SVMs have been widely applied, often paired with features like TF-IDF or linguistic markers, across tasks such as classifying code-mixed tweets [95], news headlines [153], Yelp reviews [154], satirical news [28], and Spanish tweets [17]. However, their reliance on explicit feature engineering limits their ability to capture the complex semantics and context inherent in many forms of humor.

Decision Trees Decision Trees represent a non-linear supervised learning technique used for both classification and regression [155]. They work by recursively partitioning the dataset into smaller subsets based on the values of input features. At each node of the tree, a test is applied to a specific feature (e.g., presence of a certain word, n-gram frequency, semantic feature value), and the outcome determines which branch to follow. This process continues until a leaf node is reached, which represents a class label (e.g., humorous or non-humorous) or a predicted value. The feature and split point at each node are typically chosen to maximize a criterion like information gain or Gini impurity, effectively creating a hierarchical set of rules. Decision Trees have been applied extensively in humor detection, especially for tasks like irony detection in tweets [9, 81] and general humor classification [118]. Studies suggest they tend to rely on content-based features [4, 118]. While interpretable, single decision trees can be prone to overfitting and instability; their dependence on predefined splits can limit adaptability, leading naturally to the development of ensemble methods like Random Forests [1].

Naive Bayes Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features [156]. Given a class variable C (e.g., humorous/non-humorous) and a feature vector $X = (x_1, ..., x_n)$ (e.g., word counts or TF-IDF values), Naive Bayes calculates the posterior probability $P(C|X)$ for each class using Bayes' theorem: $P(C|X) = \frac{P(X|C)P(C)}{P(X)}$. The naive assumption simplifies $P(X|C)$ to the product of individual conditional probabilities $P(x_i|C)$ for each feature x_i, assuming they are independent given the class. Despite this often unrealistic assumption, Naive Bayes models are computationally efficient and have performed surprisingly well in many text classification tasks. They have been applied to humor recognition [4], Spanish tweets [157], general tweet analysis [9], punchline detection [158], pun detection [159], and Dutch humor [160]. Their performance is moderate, particularly effective when combined with appropriate text features like TF-IDF [160, 161].

Other traditional algorithms, such as Logistic Regression [79, 159] and K-Nearest Neighbors (kNN) [159], have also been explored but are generally less common in recent humor detection literature compared to SVMs, DT ensembles, and Naive Bayes. A selection of representative performance metrics for these classical classifiers is presented in Table 3.

5.2.2 Ensemble Methods

Ensemble methods combine multiple learning algorithms to improve robustness and predictive accuracy.

Random Forest (RF) RF aggregates predictions from multiple decision trees trained on different data subsets, reducing overfitting [168]. It has been used for humor detection in diverse contexts like Yelp reviews [117], tweets [162], headlines/one-liners [163], and pun detection [159], often improving robustness on noisy data [78].

Gradient Boosting This method builds models sequentially, with each new model correcting errors made by previous ones [169, 170]. Variants like XGBoost [165] offer optimizations. Gradient Boosting has been applied to forum posts [115], news headlines [153], SemEval tasks [166], and using sentence embeddings [118]. While powerful, it can be computationally intensive and sensitive to hyperparameters [171].

Performance metrics for these ensemble methods are summarized in Table 4.

5.3 Deep Learning architectures

Deep learning approaches automated feature extraction and improved the modeling of complex linguistic patterns, overcoming some limitations of classical methods.

Table 3: Performance Metrics for Classical Classifiers

Reference	Algorithm	Dataset(s)	Accuracy	P	R	F1
Mahajan & Zaveri (2024)[154]	SVM	Yelp Reviews	-	-	-	83.32%
Jaiswal et al. (2019)[9]	Decision Trees	Tweets	81.9%	78.6%	87.6%	-
===	Naive Bayes	Tweets	81.8%	81.0%	83.5%	-
===	SVM	Tweets	86.7%	87.2%	88.3%	-
Annamoradnejad & Zoghi (2020)[118]	Decision Tree	200K Short Texts	78.6%	76.9%	82.1%	79.4%
===	SVM	===	87.2%	86.9%	88.8%	87.4%
===	Multinomial NB	===	87.6%	86.3%	90.2%	88.2%
Oliveira & Rodrigo (2019)[117]	SVM	Yelp Dataset Challenge	71.2%	-	-	-
Mihalcea & Strepparava (2006)[4]	Naive Bayes	Online Proverb Collection	84.81%	-	-	-
===	SVM	===	96.09%	-	-	-
Castro et al. (2016)[17]	SVM	Spanish Tweets	92.5%	68.9%	83.6%	75.5%
Winters & Delobelle (2020)[160]	Naive Bayes	Dutch Joke Dataset	51%	-	-	49.3%
Prajapati et al. (2023)[162]	Logistic Regression	8000 Tweets	84%	83%	83%	84%
Inacio et al. (2023)[163]	SVM	One liners (Content Feat.)[164]	-	-	-	96.4%
Fahim et al. (2024)[165]	Logistic Regression	Kaggle Humor Detection	87%	88%	87%	87%
===	SVM	===	87%	88%	88%	88%
===	Multinomial NB	===	87%	86%	85%	87%
Chi & Chi (2021)[166]	SVM	SemEval2021 Task 7 train set	55%	-	-	55%
Kumar et al. (2022)[167]	SVM	Yelp user review dataset	-	73.6%	72.8%	73.2%

Note: P: Precision, R: Recall, F1: F1-Score. '-' indicates value not reported.
"===" indicates a repeated reference or dataset from the row above.

5.3.1 Recurrent Neural Networks (RNNs)

RNNs, including variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), process sequential data by maintaining internal states, allowing them to capture temporal dependencies crucial for understanding jokes or conversational humor [172–174]. LSTMs and BiLSTMs have been used for humor prediction in dialogues [175], general text [11, 176], multilingual irony detection [119], and combined with CNNs [167]. GRUs have been applied to TWSS jokes [177], and Bi-GRUs to headlines [178]. RNNs have also been used for audio-based humor prediction [179].

5.3.2 Convolutional Neural Networks (CNNs)

Originally from image processing, CNNs were adapted for text by using filters to extract local patterns (like n-grams) from word or character sequences [180, 181]. They capture hierarchical features relevant to humor, such as specific word combinations or stylistic elements. CNNs have been applied to humor recognition in TED Talks [182], pun datasets [182, 183], Chinese humor [183], predicting audience laughter [184], and in hybrid CNN-LSTM models for Yelp reviews [167].

Performance metrics for these early deep learning methods are included in Table 5.

5.4 Performance Analysis and Discussion

The progression from classical machine learning to ensemble methods and early deep learning architectures marks a clear trend of improving performance in humor detection, as seen in the selected results in Table **??**. Classical methods established baselines, with F1 scores often ranging from roughly 50% to the high 80s, peaking occasionally over 90% on specific datasets like proverbs or one-liners using SVM [4, 163]. Performance heavily depended on feature engineering and dataset characteristics. Key datasets from this era show representative performance: Yelp Reviews (F1 $\sim$83%) [154], Tweets (Acc $\sim$87%) [9], 200K Short Texts (F1 $\sim$88%) [118], and Spanish Tweets (F1 $\sim$75%) [17].

Table 4: Performance Metrics for Ensemble Methods

Reference	Algorithm	Dataset(s)	Accuracy	P	R	F1
Annamoradnejad & Zoghi (2020)[118]	XGBoost	200K Short Texts	72.0%	75.3%	77.7%	81.3%
Yang et al. (2015)[1]	Random Forest	Pun of the Day	85.4%	83.4%	88.8%	85.9%
===	===	16000 One Liners	79.7%	77.6%	83.6%	80.5%
Oliveira & Rodrigo (2015)[117]	Random Forest	Yelp Dataset Challenge	72.02%	-	-	-
Prajapati et al. (2023)[162]	Random Forest	8000 Tweets	82%	82%	82%	82%
Inacio et al. (2023)[163]	Random Forest	One Liners & Headlines (All Feat.)[164]	-	-	-	97.1%
Fahim et al. (2024)[165]	Random Forest	Kaggle Humor Detection	83%	85%	85%	83%
===	XGBoost	===	80%	79%	85%	82%
===	Ensemble (LR, k-Means, GB)	===	93%	94%	93%	93%
Chi & Chi (2021)[166]	XGBoost	SemEval2021 Task 7 train set	52%	-	-	52%
===	Random Forest	===	53%	-	-	52%
Kumar et al. (2022)[167]	Random Forest	Yelp user review dataset	-	74.1%	73.6%	73.8%
===	XGBoost	===	-	76.3%	75.1%	75.9%

Note: P: Precision, R: Recall, F1: F1-Score. '-' indicates value not reported.
"===" indicates a repeated reference or dataset from the row above.

Table 5: Performance Metrics for Early Deep Learning Architectures

Reference	Algorithm	Dataset(s)	Accuracy	P	R	F1
Annamoradnejad & Zoghi (2020)[118]	XLNet	200K Short Texts	91.6%	87.2%	97.3%	92.0%
Winters & Delobelle (2020)[160]	LSTM	Dutch Joke Dataset	94.0%	-	-	94.0%
Prajapati et al. (2023)[162]	CNN	8000 Tweets	83.3%	83.4%	83.2%	83.3%
===	LSTM	===	87.0%	86.5%	86.1%	86.3%
Patel et al. (2021)[11]	LSTM	200K Short Texts	94.62%	-	-	-
Kumar et al. (2022)[167]	CNN-LSTM	Yelp user review dataset	-	87.6%	87.2%	87.4%
Bertero & Fung (2020)[179]	CNN	The Big Bang Theory	73.8%	70.3%	66.7%	68.5%
Tasnia et al. (2022)[119]	LSTM	SemEval-2021 Task 7	93.8%	92.2%	93.5%	93.6%
Chen & Lee (2017)[182]	CNN	Pun of the Day Corpus	86.4%	86.1%	86.4%	85.7%
Chen & Soo (2018)[183]	CNN	Pun of the Day	89.4%	86.6%	94%	90.1%
===	===	PTT Jokes	92.7%	95.7%	95.9%	94.3%
===	===	16000 One Liners	89.7%	87.2%	93.6%	90.3%

Note: P: Precision, R: Recall, F1: F1-Score. '-' indicates value not reported.
"===" indicates a repeated reference or dataset from the row above.

Ensemble methods generally offered incremental improvements or increased robustness, particularly on noisy social media data (e.g., Tweets, Acc ∼90%) [9]. While achieving high scores (up to 97.1% F1 on One Liners & Headlines) [163], gains over the best classical models on shared datasets like Yelp (Acc ∼72%) [117] or SemEval Task 7 (F1 ∼52-53%) [166] were sometimes marginal. XGBoost performed well on 200K Short Texts (F1 81.3%) [118] and Kaggle Humor (F1 82%) [165].

Early deep learning methods, such as RNNs, LSTMs, and CNNs, demonstrated more significant advances by automating feature learning and better capturing sequential context. On datasets like Yelp reviews, DL models reached ∼87% F1 compared to ∼74-83% for classical/ensemble methods [167]. On 200K Short Texts, DL (XLNet, LSTM) achieved 92-94% F1/Accuracy, surpassing the ∼81-88% range of previous methods [11, 118]. The most dramatic improvement was seen on SemEval Task 7, jumping from ∼52-55% F1 to ∼93.6% F1 [119, 166]. Pun of the Day also saw strong CNN performance (F1 ∼90%) [183]. These trends highlight the increasing effectiveness of leveraging learned representations and sequential modeling, setting the stage for the transformer-based LLMs discussed next.

6 Large Language Models for Humor Detection

6.1 The Transformer Revolution: Foundations of LLM-Based Humor Detection

The introduction of the transformer architecture [185] has fundamentally transformed natural language processing, enabling substantial advances in tasks ranging from machine translation to sentiment analysis[186]. Through successive generations of development, transformer-based models have scaled from millions to trillions of parameters, exemplified by contemporary systems such as GPT-4o[187] and Llama 3.1[188] that demonstrate unprecedented capabilities in contextual processing. This subsection examines the architectural foundations, pre-training methodologies, multilingual extensions, and scaling principles that underpin modern large language models (LLMs). We contextualize these developments within the specific domain of humor detection, illustrating how architectural innovations address longstanding limitations in sequential modeling while simultaneously introducing novel computational and interpretability challenges.

6.1.1 The Paradigm Shift from Traditional Methods to Transformers

The transformer architecture, introduced by Vaswani et al. [185], marked a significant shift from earlier recurrent (RNN, LSTM) and convolutional networks used in NLP. By employing self-attention mechanisms, transformers process sequences in parallel, effectively mitigating the vanishing gradient and long-dependency limitations inherent in sequential models [172, 173, 189]. This parallel processing capability drastically accelerated training and enabled models to capture global relationships within text.

This architectural advance paved the way for large language models (LLMs) with diverse specializations, such as bidirectional encoders like BERT [190] for classification tasks and autoregressive decoders like GPT [191] for generation. Transformers have demonstrated substantial gains over RNNs in various NLP tasks [192], and their ability to model context and leverage commonsense knowledge has been applied to humor generation [193] and explanation [194]. However, this paradigm necessitates large datasets for pre-training [195] and raises challenges regarding interpretability and applicability to low-resource scenarios. Nonetheless, the core mechanisms provide a flexible base for extensions like multimodal fusion [196], establishing the foundation for modern LLM-based approaches to complex tasks like humor detection.

6.1.2 Core Transformer Architecture

Transformers consist of stacked encoder and/or decoder blocks [185]. Input tokens are first embedded into continuous vectors and combined with positional encodings to retain sequence information. The central component is the self-attention mechanism, which calculates the contextual relevance between different tokens in the sequence, allowing the model to weigh the importance of different words when representing a specific word. Multi-head attention performs this process multiple times in parallel with different learned transformations, capturing diverse contextual relationships.

Each block also contains position-wise feed-forward networks (FFNs) for further processing, along with residual connections and layer normalization to stabilize training and enable deeper architectures. Figure 1 provides a visual overview. Architectural variants exist: encoder-only models (e.g., BERT [190]) are suited for analysis tasks, decoder-only models (e.g., GPT [191]) excel at generation, and encoder-decoder models (e.g., T5 [197]) handle sequence-to-sequence tasks. Optimizations like FlashAttention [198] have improved efficiency for long sequences. While powerful for modeling dependencies, the complexity of attention mechanisms can make models susceptible to overfitting on noisy data, a relevant concern for the subtleties and ambiguities present in humor detection.

6.2 LLM Adaptation Strategies for Humor Detection

Recent advances in large language models (LLMs) have motivated the exploration of various adaptation strategies to tailor general-purpose models to humor detection tasks. These strategies can be broadly categorized into fine-tuning, parameter-efficient approaches, prompt engineering, and multi-task learning. This section synthesizes key trends, methodological innovations, and empirical findings across these adaptation paradigms.

6.2.1 Fine-Tuning Approaches

Fine-tuning remains the dominant approach for adapting pre-trained transformers to humor understanding and generation. It involves supervised training on labeled humor datasets to update all or selected parameters, optimizing task-specific objectives such as cross-entropy loss for classification [199].

Several works have leveraged fine-tuning for humor detection across languages and modalities. For instance, Inácio et al. [200] fine-tuned BERTimbau for Portuguese humor recognition, while Gupta et al. [201] evaluated multiple

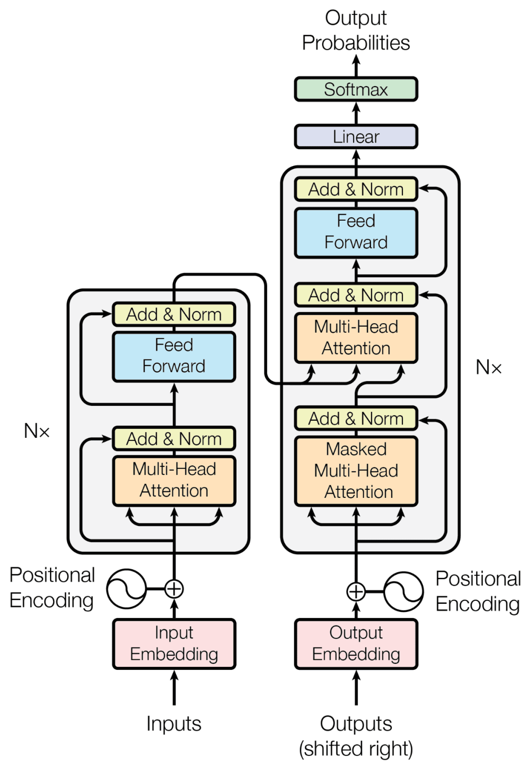

Figure 1: The Transformer model architecture. Adapted from Vaswani et al. [185].

LLMs (BERT, RoBERTa, XLNet, ERNIE 2.0, DeBERTa) with task-specific classification or regression heads. Chen et al. [202] further compared pre-trained models (BERT, RoBERTa, BART, T5, CPT) in both zero-shot and fine-tuning settings, demonstrating that fine-tuning consistently outperforms zero-shot inference.

Knowledge-augmented fine-tuning has also been explored. Chen et al. [202] integrated Pinyin embeddings into PLMs via implicit addition and explicit fusion, finding that explicit fusion notably enhanced humor-type classification. Wu et al. [203] fine-tuned LLaMA 3–8B and RoBERTa to classify humor into six genres, while Chen et al. [204] fine-tuned T5, BART, and CPT on the TalkFunny dataset to improve humorous response generation. Horvitz et al. [205] and Zhang et al. [206] extended fine-tuning to alignment settings, employing supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO) for humor generation and "unfunny" text detection.

6.2.2 Parameter-Efficient Methods

While full-model fine-tuning dominates early studies, recent works explore lightweight alternatives such as LoRA and prefix-tuning to improve efficiency. Parameter-efficient fine-tuning (PEFT) has emerged as a computationally viable alternative to full-model fine-tuning, reducing resource requirements while maintaining performance [207]. Techniques such as Low-Rank Adaptation (LoRA), adapters, and prefix-tuning introduce small trainable modules or task-specific vectors while keeping base parameters frozen.

Recent work has applied these methods to humor detection under constrained environments. Wu et al. [203] fine-tuned LLaMA 3 using 4-bit quantized LoRA (QLoRA), significantly lowering GPU memory usage without major accuracy loss. Similarly, Horvitz et al. [205] adopted QLoRA to fine-tune a Mistral classifier on humor-related tasks, showing that quantized PEFT remains effective for stylistically sensitive domains like humor.

6.2.3 Prompt Engineering Techniques

Prompt engineering leverages LLMs' in-context learning abilities to perform humor-related tasks without parameter updates [208]. This paradigm encompasses zero-shot prompting [209], few-shot prompting with exemplars [210], and chain-of-thought (CoT) prompting [211] for stepwise reasoning.

Bago et al. [212] employed few-shot prompting with GPT-4 and Gemini 1.5 Flash for Croatian humor detection, embedding four fixed examples in English prompts that required Croatian rationales. Horvitz et al. [205] used few-shot prompts to guide GPT-4 and Mistral in editing humorous versus non-humorous text ("unfunning"). Chen et al. [204] found that explicit instruction prompts ("Please answer with a touch of humor") significantly improved GPT-3.5's humor generation.

Comparative analyses suggest mixed outcomes for CoT prompting. He et al. [16] reported that CoT often degraded performance relative to direct prompting (DP) in humor explanation tasks involving GPT-4o and ERNIE4-turbo. Wu et al. [203] similarly observed limited benefits of CoT prompting when fine-tuning LLaMA models on the JOKER shared task dataset.

6.2.4 Multi-Task Learning Approaches

Multi-task learning (MTL) enables shared representations across humor-related subtasks—such as detection, rating, sentiment, and style classification—enhancing model generalization. Gupta et al. [201] trained a single transformer jointly on four humor subtasks using hard parameter sharing, observing improved regression task performance. Chen et al. [204] extended this idea by jointly training humor generation with auxiliary tasks (e.g., sentiment-style classification, rewriting), demonstrating synergistic learning effects. Together, these studies indicate that MTL mitigates data scarcity and overfitting in humor datasets.

6.3 Multi-Modal and Cross-Lingual Extensions

6.3.1 Vision-Language Models for Visual Humor

Vision-language models (VLMs) extend humor understanding beyond text. Zhang et al. [206] evaluated models like LLaVA and GPT-4o Vision on humorous cartoon captioning. Their findings show that current multimodal models underperform text-only LLMs, suggesting that visual humor comprehension remains an open challenge requiring improved cross-modal grounding.

6.3.2 Cross-Lingual Humor Detection

Humor's cultural dependency motivates multilingual and cross-lingual investigations. He et al. [16] evaluated LLMs on Chinese humor from the Ruo Zhi Ba (RZB) platform, highlighting difficulties in handling tonal and character-based ambiguity. Chen et al. [202] constructed a large-scale Chinese humor dataset to address linguistic and phonetic humor phenomena such as homophony and pun-based play. Bago et al. [212] studied humor detection in Croatian through bilingual prompting, while Horvitz et al. [205] examined code-mixed English–Hindi humor, using GPT-4-generated data to train XLM-RoBERTa classifiers. Collectively, these works underline the importance of cultural context and multilingual grounding for generalizable humor modeling.

6.4 Evaluation, Benchmarks, and Performance Analysis

6.4.1 Standard Evaluation Metrics

LLM-based humor detection typically employs metrics such as Accuracy [16, 201–203], Precision, Recall, and F1-score [213], with Macro-F1 used for imbalanced datasets [212]. Additional metrics include Mean Average Precision/Recall [203], False Positive/Negative Rates [16], Matthews Correlation Coefficient (MCC) [16], Root Mean Squared Error (RMSE) for regression [201], and Cohen's Kappa for annotation consistency [212]. Recent works also use A/B testing win rates to compare human and model-generated humor explanations.

6.4.2 Benchmark Datasets

Numerous datasets underpin contemporary humor detection research. The JOKER shared task [3, 214] and SemEval datasets [13, 122, 153] remain key English-language benchmarks. TalkFunny [204] and Chumor [16] provide large Chinese humor datasets emphasizing explanation and reasoning. HRumor 1.0 [212] introduces Croatian humor annotations, while Chen et al. [202] offer a Chinese multi-task humor suite. Other notable datasets include ColBERT [118], URFUNNY [131], M2H2 [133], and PixelHumor [135]. For multimodal benchmarks, MHSDB [215] standardizes humor and sarcasm detection across modalities, and LS-FUNNY [137] introduces a multilingual audiovisual corpus from Spanish TEDx talks. Large-scale web collections such as Reddit r/Jokes [216] and the FUN Russian corpus [217] expand linguistic and cultural diversity.

Table 6: Detailed Performance Metrics for Transformer Architectures

Reference	Algorithm	Dataset(s)	Accuracy	P	R	F1
Winters & Delobelle (2020)[160]	RobBERT	Dutch Joke Dataset	98.8%	-	-	98.8%
Inacio et al. (2023)[163]	Finetuned BERT	One liners and Headlines (All Features)[164]	-	-	-	99.6%
Annamoradnejad & Zoghi (2020)[118]	ColBERT	200K Short Texts	98.2%	99%	97.4%	98.2%
Inacio et al. (2024)[200]	BERTimbau-large	Portuguese Jokes	-	-	-	68.7%
===	BERTimbau-base	===	-	-	-	67.8%
===	Albertina-900M PT-PT	===	-	-	-	52.1%
===	Albertina-900M PT-BR	===	-	-	-	51.5%
Guo et al. (2022)[213]	FedHumor	SemEval-2020 Shared Task 7	-	66.6%	66.56%	66.53%
Wu et al. (2025)[203]	Llama 3-8B (with SFT)	JOKER Dataset	69.78%	-	-	-
===	RoBERTa	===	68.14%	-	-	-
Gupta et al. (2021)[201]	RoBERTa	SemEval-2021 Task 7 Dataset	94.1%	-	-	95.2%
===	DeBERTa	===	94.2%	-	-	95.3%

Note: P: Precision, R: Recall, F1: F1-Score.
"===" indicates a repeated entry from the row above.

6.4.3 Performance Comparison and Human-Model Agreement

Comparative evaluations reveal wide variance in LLM performance across datasets and prompting settings. Gupta et al. [201] reported high accuracy (94–95%) on SemEval tasks using RoBERTa and DeBERTa, while Wu et al. [203] found that even advanced instruction-tuned LLMs (e.g., LLaMA 3–8B, DeepSeek-R1, Qwen2.5) underperform humans on JOKER datasets (14–21% accuracy). He et al. [16] observed substantial human–model gaps: human annotators achieved 78.3% accuracy and MCC 0.60 on humor explanation classification, whereas the best LLM (ERNIE4-turbo) reached 60.3% accuracy and MCC 0.29. A/B tests confirmed human explanations were overwhelmingly preferred. These results collectively highlight that while LLMs capture surface humor cues, they still lack deeper pragmatic and cultural reasoning.

6.5 Comprehensive Performance Comparison

Understanding the comparative performance of transformer-based and large language models (LLMs) in humor detection tasks provides critical insights into the architectural trade-offs, fine-tuning strategies, and linguistic adaptability that define state-of-the-art models. Transformer variants such as BERT, RoBERTa, and ColBERT have consistently demonstrated strong performance across diverse datasets, while recent LLMs like GPT-4, ERNIE4-turbo, and Deepseek-R1 reveal the challenges of scaling humor understanding across multilingual and culturally nuanced contexts.

Transformer Architectures

Table 6 summarizes key transformer-based architectures applied to humor detection. Models such as RobBERT [160] and ColBERT [118] achieve near-human accuracy on monolingual datasets, underscoring the value of contextual embeddings and fine-tuning strategies. Meanwhile, multilingual variants such as BERTimbau and Albertina show varying degrees of transfer performance, highlighting the persistent gap in cross-lingual generalization and cultural humor comprehension.

LLM Performance Comparison

Beyond classical transformer-based models, large-scale instruction-tuned LLMs exhibit different behaviors on humor detection tasks, particularly under different reasoning paradigms such as **Direct Prediction (DP)** versus **Chain-of-Thought (CoT)** prompting. He et al. [16] benchmarked ERNIE4-turbo, Gemini 1.5 Pro, and GPT-4 variants on CHUMOR 2.0, illustrating that CoT prompting does not consistently outperform DP in humor understanding—a sign that current LLMs lack robust pragmatic grounding. Conversely, Wu et al. [203] found that parameter scaling alone (e.g., Deepseek-R1:671B vs. 32B) offers limited benefit, indicating the importance of fine-tuning and domain adaptation.

Table 7: Detailed Performance Metrics for Large Language Models on Humor Detection

Reference	**Algorithm**	**Dataset(s)**	**Accuracy**	**P**	**R**	**F1**
Wu et al. (2025)[203]	Deepseek-R1:671B	JOKER Lab 2024 Shared Task	21.08%	-	-	-
===	Deepseek-R1:32B-q4	===	17.07%	-	-	-
===	Qwen2.5:7B	===	16.32%	-	-	-
===	Llama 3-8B (without SFT)	===	14.93%	-	-	-
===	QwQ:32B	===	14.13%	-	-	-
He et al. (2025)[16]	ERNIE4-turbo (DP)	CHUMOR 2.0	60.3%	-	-	-
===	ERNIE4-turbo (CoT)	===	45.2%	-	-	-
He et al. (2025)[16]	Gemini 1.5 Pro (DP)	CHUMOR 2.0	54.0%	-	-	-
===	Gemini 1.5 Pro (CoT)	===	60.3%	-	-	-
He et al. (2025)[16]	QWen2.572B (DP)	CHUMOR 2.0	48.5%	-	-	-
===	QWen2.572B (CoT)	===	49.5%	-	-	-
He et al. (2025)[16]	GPT-4-turbo (DP)	CHUMOR 2.0	52.3%	-	-	-
===	GPT-4-turbo (CoT)	===	51.3%	-	-	-
He et al. (2025)[16]	GPT-4o (DP)	CHUMOR 2.0	51.9%	-	-	-
===	GPT-4o (CoT)	===	50.6%	-	-	-
Gupta et al. (2021)[201]	ERNIE-2.0	SemEval-2021 Task 7 Dataset	94.3%	-	-	95.4%

Note: P: Precision, R: Recall, F1: F1-Score.
"===" indicates a repeated entry from the row above.

Discussion. Across both transformer-based and LLM paradigms, results reveal a consistent tension between task-specific fine-tuning and general reasoning capabilities. Smaller, domain-adapted transformers outperform massive general-purpose LLMs in humor detection, particularly in languages beyond English. This underscores humor's strong dependency on cultural, linguistic, and pragmatic context—dimensions that remain underrepresented in current pretraining corpora. Despite scaling, reasoning-oriented LLMs still struggle to model irony, satire, and cultural reference without additional grounding or alignment to human humor norms.

6.5.1 Cross-Dataset Generalization

Cross-dataset transfer remains limited due to heterogeneous annotation schemes and label granularity. Gupta et al. [201] observed that masked language model–based augmentation degraded humor semantics, underscoring the sensitivity of humor meaning to linguistic context. Broader efforts toward unified benchmarks and multilingual humor taxonomies are needed to enable consistent evaluation and transfer across datasets.

6.6 Explainability and Interpretability in LLM Humor detection

Understanding *why* an LLM classifies text as humorous remains a significant challenge, intertwined with the broader difficulty of interpreting complex neural models and the inherent subjectivity of humor itself. While LLMs demonstrate increasing capabilities in detecting surface patterns associated with humor, their internal reasoning processes often remain opaque.

Attempts to enhance interpretability in humor tasks have explored several avenues:

- **Generating Explanations:** Some research explicitly prompts LLMs to generate explanations for their humor judgments or for why a piece of text is funny [16, 135]. He et al. [16] created the Chumor dataset specifically for evaluating the quality of such explanations, finding through human A/B testing that LLM-generated explanations (from GPT-4o, ERNIE4-turbo) are still significantly less preferred than human-written ones, often relying on generic reasoning patterns about absurdity or unexpectedness. Ryan et al. [135] similarly included a humor interpretation task in their PixelHumor benchmark, evaluating explanation relevance via human ratings.
- **Chain-of-Thought (CoT) Prompting:** While CoT prompting aims to make reasoning steps explicit [211], its application to humor has yielded mixed results. He et al. [16] found that CoT often degraded performance in classifying explanation quality compared to direct prompting. Wu et al. [203] also noted limited benefits during fine-tuning. This suggests that the complex, often non-linear reasoning involved in humor may not align well with current CoT elicitation methods.
- **Probing and Analysis:** Other works implicitly touch upon interpretability by analyzing model failures or comparing performance across different humor types or contexts [202, 206]. For example, identifying specific humor types (like sarcasm or culturally nuanced jokes) that models struggle with provides indirect insight into their limitations [16, 135].

Overall, explainability in LLM-based humor detection is nascent. Current methods primarily focus on generating post-hoc textual justifications, whose faithfulness to the model's actual decision process is uncertain. Developing techniques that provide deeper insights into how LLMs process incongruity, cultural references, and pragmatic cues remains a critical area for future research, essential for building trust and diagnosing failures in conversational AI systems.

7 Open Challenges, Limitations and Future Directions

Despite significant progress driven by LLMs, automatic humor detection faces numerous open challenges and limitations, paving the way for future research directions.

- **Subjectivity and Cultural Nuance:** Humor remains deeply subjective and culturally dependent [16, 202]. LLMs trained predominantly on Western, English-centric data struggle with culturally specific references, wordplay, and pragmatic norms prevalent in other languages and communities [16, 212]. Developing culturally aware and adaptable models is paramount.
- **Multimodal Understanding:** While VLMs are emerging, effectively integrating visual, auditory, and textual cues for humor detection remains challenging [206, 215]. Current models often underperform text-only counterparts or fail to capture humor arising from cross-modal interactions (e.g., timing in video, visual puns in comics) [135, 206]. Future work needs better cross-modal fusion, grounding, and temporal reasoning.
- **Data Scarcity and Quality:** High-quality, large-scale, and diverse annotated datasets are scarce, especially for non-English languages and multimodal humor [137, 212]. Existing datasets often suffer from annotation inconsistencies due to humor's subjectivity [135] or rely on distant supervision (e.g., laughter tracks, hashtags) which introduces noise [137]. Curating richer, more reliable, and culturally diverse benchmarks is crucial.
- **Deep Reasoning vs. Surface Cues:** LLMs excel at capturing surface linguistic patterns but often lack the deep pragmatic, contextual, and world knowledge required for complex humor like satire, irony, or subtle incongruity [16, 203]. Bridging this gap requires models capable of more sophisticated inference and common-sense reasoning.
- **Evaluation and Benchmarking:** Current evaluation often relies on standard classification metrics that may not fully capture the nuances of humor understanding [201]. Heterogeneity in datasets and tasks limits cross-study comparisons [215]. There is a need for unified benchmarks, potentially incorporating human-in-the-loop evaluation or metrics sensitive to humor type and quality [16, 206].
- **Adaptation Strategies:** The effectiveness of different adaptation strategies (fine-tuning, PEFT, prompting, alignment) for creative and subjective tasks like humor is still under-explored. For instance, RLHF appears sensitive in humor domains [206], and CoT shows mixed results [16, 203]. Research is needed to develop robust adaptation methods tailored for subjective language understanding.

- **Humor vs. Offensiveness:** Humor often borders on sensitive topics, and distinguishing benign violations from genuinely offensive content is a critical challenge [201]. Models must be developed with safeguards to avoid generating or misinterpreting harmful humor, requiring careful dataset curation and alignment strategies [206].
- **Explainability:** As highlighted in the previous section, understanding and explaining *why* an LLM perceives something as humorous remains a major hurdle, limiting trust and diagnostic capabilities.

Future directions should focus on creating more diverse and robust datasets, developing culturally grounded and multimodally adept models, refining adaptation and alignment techniques for subjective tasks, establishing more meaningful evaluation protocols, and advancing explainability methods to demystify computational humor reasoning.

8 Conclusion

Automatic humor detection, a long-standing challenge at the intersection of AI, linguistics, and psychology, has witnessed significant evolution, transitioning from theory-driven feature engineering and classical machine learning to sophisticated deep learning and, most recently, large language models (LLMs). This survey has traced this trajectory, highlighting the theoretical underpinnings, computational methodologies, benchmark datasets, and evaluation paradigms that have shaped the field.

LLMs, with their unprecedented scale and emergent capabilities, offer new potential for capturing the complex linguistic, contextual, and cultural nuances inherent in humor. Adaptation strategies like fine-tuning, parameter efficient fine-tuning, and prompt engineering allow these general-purpose models to be specialized for humor-related tasks, while multimodal extensions are beginning to tackle humor beyond pure text. Performance on certain benchmarks, particularly those involving pattern recognition or explicit humor markers, has improved considerably.

However, significant challenges persist. LLMs still demonstrate limitations in deep pragmatic reasoning, cultural understanding, and effective multimodal integration necessary for robust humor comprehension that mirrors human ability. Performance gaps between models and human judgment remain substantial, especially for subtle, ironic, or culturally specific humor. Furthermore, issues surrounding data scarcity, evaluation consistency, model explainability, and the delicate balance between humor and offensiveness continue to hinder progress.

The pursuit of computational humor detection is not merely an academic exercise; it is crucial for developing AI systems that can interact naturally, engagingly, and appropriately with humans. As LLMs become more deeply embedded in our daily lives, their ability to understand and navigate the complexities of human social expression, including humor, will be essential for fostering effective and trustworthy human-AI collaboration. Continued interdisciplinary research, focusing on culturally grounded models, richer multimodal datasets, robust evaluation frameworks, and transparent reasoning processes, will be key to unlocking the next generation of socially intelligent AI.

References

[1] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] Tanisha Khurana, Kaushik Pillalamarri, Vikram Pande, and Munindar Singh. LOLgorithm: Integrating Semantic,Syntactic and Contextual Elements for Humor Classification, August 2024. arXiv:2408.06335 [cs].

[3] Liana Ermakova, Anne-Gwenn Bosser, Tristan Miller, Tremaine Thomas, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. CLEF 2024 JOKER Lab: Automatic Humour Analysis. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, pages 36–43, Cham, 2024. Springer Nature Switzerland.

[4] Rada Mihalcea and Carlo Strapparava. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142, 2006.

[5] Cristina Larkin-Galiñanes. An overview of humor theory. 2017.

[6] Tabea Scheel. *Definitions, Theories, and Measurement of Humor*, pages 9–29. 09 2017.

[7] María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Miguel Ángel Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33, July 2017.

[8] Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. Explaining humour style classifications: An xai approach to understanding computational humour analysis. *Journal of Data Mining amp; Digital Humanities*, NLP4DH(Digital humanities in...), April 2025.

[9] Arunima Jaiswal, Monika, Anshu Mathur, Prachi, and Sheena Mattu. Automatic humour detection in tweets using soft computing paradigms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 172–176, 2019.

[10] Tina Esther Trueman, Gopi K, and Ashok Kumar J. Online Text-Based Humor Detection. In *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, volume 1, pages 313–316, November 2021.

[11] Krupa Patel, Manasi Mathkar, Sarjak Maniar, Avi Mehta, and Prof. Shachi Natu. To laugh or not to laugh – LSTM based humor detection approach. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7, July 2021.

[12] Issa Annamoradnejad and Gohar Zoghi. ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685, September 2024.

[13] Fara Shatnawi, Malak Abdullah, and Mahmoud Hammad. MLEngineer at SemEval-2020 Task 7: BERT-Flair Based Humor Detection Model (BFHumor). In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1041–1048, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[14] Kexin Quan, Pavithra Ramakrishnan, and Jessie Chin. Can AI Take a Joke—Or Make One? A Study of Humor Generation and Recognition in LLMs. In *Proceedings of the 2025 Conference on Creativity and Cognition*, C&C '25, pages 431–437, New York, NY, USA, June 2025. Association for Computing Machinery.

[15] Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. From Punchlines to Predictions: A Metric to Assess LLM Performance in Identifying Humor in Stand-Up Comedy, April 2025. arXiv:2504.09049 [cs].

[16] Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. Chumor 2.0: Towards Benchmarking Chinese Humor Understanding, December 2024. arXiv:2412.17729 [cs].

[17] Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. Is this a joke? detecting humor in spanish tweets. volume 10022, pages 139–150, 11 2016.

[18] Rida Miraj and Masaki Aono. Humor detection in spanish tweets using neural network. In *IberLEF@ SEPLN*, pages 837–843, 2021.

[19] Valentin Barriere, Nahuel Gomez, Leo Hemamou, Sofia Callejas, and Brian Ravenet. StandUp4AI: A New Multilingual Dataset for Humor Detection in Stand-up Comedy Videos, May 2025. arXiv:2505.18903 [cs].

[20] Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T. Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms, 2025.

[21] Lukas Christ, Shahin Amiriparian, Alexander Kathan, Niklas Müller, Andreas König, and Björn W. Schuller. Towards Multimodal Prediction of Spontaneous Humor: A Novel Dataset and First Results. *IEEE Transactions on Affective Computing*, 16(2):844–860, April 2025.

[22] Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. Systematic Literature Review: Computational Approaches for Humour Style Classification, January 2024. arXiv:2402.01759 [cs].

[23] Anirudh Mittal, Diptesh Kanojia, and Pushpak Bhattacharyya. Survey on computational humour, 2022.

[24] C. Ren, Z. Guo, P. Zhang, and Y. Gao. Humor detection using deep learning in 10 years: A survey. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, 40(1):4, 2024.

[25] Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In Stefania DeGaetano, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online, December 2020. International Committee on Computational Linguistics.

[26] THOMAS C. VEATCH. A theory of humor. *HUMOR*, 11(2):161–216, 1998.

[27] Victor Raskin. *Semantic Mechanisms of Humor*, volume 5. 01 1985.

[28] Tanvir Ahmad, Halima Akhtar, Akshay Chopra, and Mohd Waris Akhtar. Satire detection from web documents using machine learning methods. In *2014 international conference on soft computing and machine intelligence*, pages 102–105. IEEE, 2014.

[29] Victor Raskin. Linguistic heuristics of humor: a script-based semantic approach. *International Journal of the Sociology of Language*, 1987(65):11–26, 1987.

[30] Arthur Asa Berger. *An Anatomy of Humor*. Routledge, 1st edition, 1993.

[31] Arthur Asa Berger. Why We Laugh and What Makes Us Laugh: 'The Enigma of Humor'. *Europe's Journal of Psychology*, 9(2):210–213, May 2013.

[32] Jackson G. Lu. Cultural differences in humor: A systematic review and critique. *Current Opinion in Psychology*, 53:101690, October 2023.

[33] Peter McGraw and Joel Warner. *The Humor Code: A Global Search for What Makes Things Funny*. Simon & Schuster, New York, 2014. Includes bibliographical references (pages 218-226) and index.

[34] T. Jiang, H. Li, and Y. Hou. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10:123, 2019. Received: 26 November 2018; Accepted: 14 January 2019; Published: 29 January 2019.

[35] ARTHUR ASA BERGER. Humor: An introduction. *American Behavioral Scientist*, 30(3):6–15, 1987.

[36] Jessalyn I. Vallade, Melanie Booth-Butterfield, and Lori E. Vela. Taking Back Power: Using Superiority Theory to Predict Humor Use Following a Relational Transgression. *Western Journal of Communication*, 77(2):231–248, March 2013.

[37] SHEILA LINTOTT. Superiority in Humor Theory. *The Journal of Aesthetics and Art Criticism*, 74(4):347–358, October 2016.

[38] Jarno Hietalahti. Laughing at oneself: On the new social character. *Studies in Social and Political Thought*, 25, December 2015.

[39] John C. Meyer. Humor as a double-edged sword: Four functions of humor in communication. *Communication Theory*, 10(3):310–331, 03 2006.

[40] Mohammad Ahmadi. Book review: Attardo, salvatore (2024). linguistic theories of humour. de gruyter mouton. *The European Journal of Humour Research*, 13:287–292, 07 2025.

[41] Arthur Shurcliff. Judged humor, arousal, and the relief theory. *Journal of personality and social psychology*, 8 4:360–3, 1968.

[42] Shane Sizemore and Kimberly O'Brien. How to tell a joke: theories of successful humor and applications to the workplace. *Management Research Review*, 46(12):1679–1693, October 2023.

[43] Tabea Scheel and Christine Gockel. *Humor at Work in Teams, Leadership, Negotiations, Learning and Health*. 09 2017.

[44] Eric Romero and Kevin Cruthirds. The use of humor in the workplace. *Engineering Management Review, IEEE*, 34:18–18, 01 2007.

[45] Christine Gockel. *Humor in Teams: Interpersonal Functions of Humor*, pages 31–46. 09 2017.

[46] Christine Gockel and Laura Vetter. *Humor in Leadership: How to Lead People with Humor*, pages 47–63. 09 2017.

[47] Tae-Yeol Kim, Deog Lee, and Noel Wong. Supervisor humor and employee outcomes: The role of social distance and affective trust in supervisor. *Journal of Business and Psychology*, 31:125–139, 02 2016.

[48] Brita Banitz. *Funny business: Verbal humor in business negotiation and the English-as-a-second-language speaker*. Ph.d. dissertation, Purdue University, West Lafayette, IN, August 2005.

[49] Christine Gockel. *Humor in Negotiations: How to Persuade Others with Humor*, pages 65–77. 09 2017.

[50] Tabea Scheel. *Humor and Learning in the Workplace*, pages 79–94. 09 2017.

[51] Simon A. Lei, Jillian Lane Cohen, and Kristen Russler. Humor on learning in the college classroom: Evaluating benefits and drawbacks from instructors' perspectives. *Journal of Instructional Psychology*, 37:326–332, 2010.

[52] Ryota Tsukawaki and Tomoya Imura. Students' perception of teachers' humor predicts their mental health. *Psychological Reports*, 125(1):98–109, 2022. PMID: 33148116.

[53] Nancy D Bell. Humor comprehension: Lessons learned from cross-cultural communication. *HUMOR*, 20(4):367–387, 2007.

[54] Yi Cao, Yubo Hou, Zhiwen Dong, and Li-Jun Ji. The impact of culture and social distance on humor appreciation, sharing, and production. *Social Psychological and Personality Science*, 14(2):207–217, 2023.

[55] Jyotsna Vaid. Joking across languages: Perspectives on humor, emotion, and bilingualism. *Bilingual Education and Bilingualism*, 56:152, 2006.

[56] Sonja Heintz, Willibald Ruch, Tracey Platt, Dandan Pang, Hugo Carretero-Dios, Alberto Dionigi, Catalina Argüello Gutiérrez, Ingrid Brdar, Dorota Brzozowska, Hsueh-Chih Chen, et al. Psychometric comparisons of benevolent and corrective humor across 22 countries: The virtue gap in humor goes international. *Frontiers in psychology*, 9:92, 2018.

[57] Julie Aitken Schermer, Radosław Rogoza, Maria Magdalena Kwiatkowska, Christopher Marcin Kowalski, Sibele Aquino, Rahkman Ardi, Henrietta Bolló, Marija Branković, Razieh Chegeni, Jan Crusius, et al. Humor styles across 28 countries. *Current Psychology*, 42(19):16304–16319, 2023.

[58] Yihui Guo. A cross-cultural study of humor intensity in chinese and english family jokes: A large language model-based approach. *International Journal of Linguistics Studies*, 5(2):01–10, Jul. 2025.

[59] Jackson G. Lu, Ashley E. Martin, Anastasia Usova, and Adam D. Galinsky. Chapter 9 - creativity and humor across cultures: Where aha meets haha. In Sarah R. Luria, John Baer, and James C. Kaufman, editors, *Creativity and Humor*, Explorations in Creativity Research, pages 183–203. Academic Press, 2019.

[60] Dana L. Alden, Wayne D. Hoyer, and Chol Lee. Identifying global and culture-specific dimensions of humor in advertising: A multinational analysis. *Journal of Marketing*, 57(2):64–75, 1993.

[61] Guo-Hai Chen and Rod A Martin. A comparison of humor styles, coping humor, and mental health between chinese and canadian university students. 2007.

[62] Kerry Mullan and Christine Béal. Introduction: Conversational humor: Forms, functions and practices across cultures. *Intercultural Pragmatics*, 15(4):451–456, 2018.

[63] Winnie Cheng. Humor in intercultural conversations. 2003.

[64] Jennifer Tehan Stanley and Jennifer R Turner. Cross-cultural perspectives on humor appreciation and function across the lifespan. In *The Palgrave Handbook of Humour Research*, pages 455–475. Springer, 2024.

[65] Iddo Tavory. The situations of culture: Humor and the limits of measurability. *Theory and Society*, 43(3):275–289, 2014.

[66] Oybek Kulmamatov. Cross-cultural analysis of humor in linguistic and socio-cultural contexts. *Nordic_Press*, 3(0003), 2024.

[67] Julie Aitken Schermer, Radosław Rogoza, Marija Branković, Oscar Oviedo-Trespalacios, Tatiana Volkodav, Truong Thi Khanh Ha, Maria Magdalena Kwiatkowska, Eva Papazova, Joonha Park, Christopher Marcin Kowalski, et al. Humor styles are related to loneliness across 15 countries. *Europe's Journal of Psychology*, 18(4):422, 2022.

[68] Lynette S Unger. The potential for using humor in global advertising. 1996.

[69] Jennings Bryant and Dolf Zillmann. Using humor to promote learning in the classroom. *Humor and children's development*, pages 49–78, 2014.

[70] Urszula Michalik and Iwona Sznicer. The use of humor in the multicultural working environment. In *Multiculturalism, multilingualism and the self: Studies in linguistics and language learning*, pages 19–32. Springer, 2017.

[71] Salvatore Attardo and Victor Raskin. Linguistics and humor theory. In *The Routledge Handbook of Language and Humor*, pages 49–63. Routledge Handbooks Online, 2017.

[72] Victor Raskin. A little metatheory: Thought on what atheory of computational humor should look like. In *AAAI Fall Symposium: Artificial Intelligence of Humor*, 2012.

[73] Salvatore Attardo. *Linguistic Theories of Humor*. De Gruyter Mouton, Berlin, Boston, 2024.

[74] Salvatore Attardo. *The General Theory of Verbal Humor*, pages 136–156. 06 2020.

[75] Julia M. Taylor. Ontology-based view of natural language meaning: the case of humor detection. *Journal of Ambient Intelligence and Humanized Computing*, 1(3):221–234, September 2010.

[76] Sergei Nirenburg and Victor Raskin. *Ontological semantics*. Mit Press, 2004.

[77] Victor Raskin. Script-based semantic and ontological semantic theories of humor. In *The Routledge handbook of language and humor*, pages 109–125. Routledge, 2017.

[78] Antonios Kalloniatis and Panagiotis Adamidis. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43, December 2024.

[79] Yftah Ziser, Elad Kravi, and David Carmel. Humor detection in product question answering systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 519–528, 2020.

[80] Junyun Liao, Chunyu Li, and Raffaele Filieri. The role of humor in management response to positive consumer reviews. *Journal of Interactive Marketing*, 57(2):323–342, 2022.

[81] Francesco Barbieri and Horacio Saggion. Automatic detection of irony and humour in twitter. In *ICCC*, pages 155–162, 2014.

[82] Vassilis Saroglou, Christelle Lacour, and Marie-Eve Demeure. Bad humor, bad marriage: Humor styles in divorced and married couples. *Europe's Journal of Psychology*, 6(3):94–121, 2010.

[83] Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. Computational models for incongruity detection in humour. In *International conference on intelligent text processing and computational linguistics*, pages 364–374. Springer, 2010.

[84] Di Cao. Self-attention on sentence snippets incongruity for humor assessment. In *Journal of Physics: Conference Series*, volume 1827, page 012072. IOP Publishing, 2021.

[85] Bruce F Katz. A neural resolution of the incongruity-resolution and incongruity theories of humour. *Connection Science*, 5(1):59–75, 1993.

[86] Penglong Huang, Xingwei Zeng, Jinta Weng, Ying Gao, Heyan Huang, and Maobin Tang. SICKNet: A Humor Detection Network Integrating Semantic Incongruity and Commonsense Knowledge. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 288–296, October 2022. ISSN: 2375-0197.

[87] Sven van den Beukel and Lora Aroyo. Homonym detection for humor recognition in short text. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 286–291, 2018.

[88] Antonio Reyes, Davide Buscaldi, and Paolo Rosso. An analysis of the impact of ambiguity on automatic humour recognition. In *International Conference on Text, Speech and Dialogue*, pages 162–169. Springer, 2009.

[89] Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Di Wu, and Kan Xu. Crga: Homographic pun detection with a contextualized-representation: Gated attention network. *Knowledge-Based Systems*, 195:105056, 2020.

[90] Haojie Xu, Weifeng Liu, Jiangwei Liu, Mingzheng Li, Yu Feng, Yasi Peng, Yunwei Shi, Xiao Sun, and Meng Wang. Hybrid multimodal fusion for humor detection. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 15–21, 2022.

[91] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12972–12980, 2021.

[92] Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Di Wu, and Kan Xu. Homographic pun location using multi-dimensional semantic relationships. *Soft Computing-A Fusion of Foundations, Methodologies & Applications*, 24(16), 2020.

[93] Salvatore Attardo, Donalee Hughes Attardo, Paul Baltes, and Marnie Jo Petray. The linear organization of jokes: analysis of two thousand texts. 1994.

[94] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, 18(4):311–332, 2009.

[95] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A corpus of english-hindi code-mixed tweets for sarcasm detection. *arXiv preprint arXiv:1805.11869*, 2018.

[96] Yufeng Diao, Liang Yang, Dongyu Zhang, Linhong Xu, Xiaochao Fan, Di Wu, and Hongfei Lin. Homographic puns recognition based on latent semantic structures. In *National CCF conference on natural language processing and Chinese computing*, pages 565–576. Springer, 2017.

[97] Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. Uo upv: Deep linguistic humor detection in spanish social media. In *Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the Spanish society for natural language processing (SEPLN 2018)*, pages 204–213, 2018.

[98] Charvi Jain, Kshitij Sawant, Mohammed Rehman, and Rajesh Kumar. Emotion detection and characterization using facial features. In *2018 3rd International Conference and Workshops on Recent Advances and innovations in Engineering (ICRAIE)*, pages 1–6. IEEE, 2018.

[99] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

[100] Mehedi Hasan Bijoy, Dejan Porjazovski, Nhan Phan, Guangpu Huang, Tamás Grósz, and Mikko Kurimo. Multimodal humor detection and social perception prediction. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*, pages 60–64, 2024.

[101] Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375, 2021.

[102] Ashraf Kamal and Muhammad Abulaish. Self-deprecating humor detection: A machine learning approach. In *International conference of the pacific association for computational linguistics*, pages 483–494. Springer, 2019.

[103] Donghai Zhang, Wei Song, Lizhen Liu, Chao Du, and Xinlei Zhao. Investigations in automatic humor recognition. In *2017 10th International symposium on computational intelligence and design (ISCID)*, volume 1, pages 272–275. IEEE, 2017.

[104] Janyce Wiebe and Rada Mihalcea. Word sense and subjectivity. In *Proceedings of the 21st International conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072, 2006.

[105] Lizhen Liu, Donghai Zhang, and Wei Song. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 586–591, 2018.

[106] Renxian Zhang and Naishi Liu. Recognizing Humor on Twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898, Shanghai China, November 2014. ACM.

[107] Eli Borodach, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Decoders Laugh as Loud as Encoders, September 2025. arXiv:2509.04779 [cs].

[108] Motoki Yatsu and Kenji Araki. Comparison of pun detection methods using japanese pun corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[109] Da Li, Rafal Rzepka, Michal Ptaszynski, and Kenji Araki. Hemos: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media. *Information Processing & Management*, 57(6):102290, 2020.

[110] Shelley Gupta, Archana Singh, and Vivek Kumar. Emoji, text, and sentiment polarity detection using natural language processing. *Information*, 14(4):222, 2023.

[111] Rutal Mahajan and Mukesh Zaveri. Humor identification using affect based content in target text. *Journal of Intelligent & Fuzzy Systems*, 39(1):697–708, 2020.

[112] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48, 1991.

[113] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 380–390. Association for Computational Linguistics, 2013.

[114] Yoav Goldberg. *Neural network methods in natural language processing*. Morgan & Claypool Publishers, 2017.

[115] Neamul Islam Fahim, Rifah Khan, Sujana Rahman, Nusrat Akter, and Mohammad Nurul Huda. Humor detection using machine learning approach. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pages 1217–1222. IEEE, 2024.

[116] Shweta Tiwari, Jyoti Prakash Singh, Sudhakar Tripathi, and Sneha Kumari. Hybrid humor classification and detection based on machine learning. In *International Conference on Computational Intelligence in Communications and Business Analytics*, pages 103–113. Springer, 2024.

[117] Luke De Oliveira and Alfredo L Rodrigo. Humor detection in yelp reviews. *Retrieved on December*, 15:2019, 2015.

[118] Issa Annamoradnejad and Gohar Zoghi. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*, 1(3), 2020.

[119] Radiathun Tasnia, Nabila Ayman, Afrin Sultana, Abu Nowshed Chy, and Masaki Aono. Exploiting stacked embeddings with lstm for multilingual humor and irony detection. *Social Network Analysis and Mining*, 13(1):43, 2023.

[120] María Carmen Aguirre-Delgado and Angel Eduardo Cadena-Bautista. Using vector embeddings and feature vectors to humor identification. In *IberLEF@ SEPLN*, 2023.

[121] Ryan Rony Dsilva. From sentence embeddings to large language models to detect and understand wordplay. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 205–214. Springer, 2024.

[122] S Kayalvizhi, D Thenmozhi, et al. Ssn_nlp at semeval-2020 task 7: Detecting funniness level using traditional learning with sentence embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 865–870, 2020.

[123] Hou Shuo, Zhang Yijia, Wang Mengyi, Lin Hongfei, and Lu Mingyu. Cefm: Clip encoded fusion model for multimodal humor recognition on memes. *Multimedia Tools and Applications*, 84(26):31429–31443, 2025.

[124] Indradev Saw, Soumya Sahoo, Prachi Priyanka, Shashwati Jha, and Arpit Anand. Hstm: Decoding humor in memes through multimodal intelligence. In *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)*, pages 1–6. IEEE, 2025.

[125] Dongyu Zhang, Heting Zhang, Xikai Liu, Hongfei Lin, and Feng Xia. Telling the Whole Story: A Manually Annotated Chinese Dataset for the Analysis of Humor in Jokes. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6402–6407, Hong Kong, China, November 2019. Association for Computational Linguistics.

[126] Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. Mumor: A multimodal dataset for humor detection in conversations. In *CCF international conference on natural language processing and Chinese computing*, pages 619–627. Springer, 2021.

[127] Yishay Raz. Automatic humor classification on twitter. In *Proceedings of the NAACL HLT 2012 student research workshop*, pages 66–70, 2012.

[128] Bobak Farzin, Piotr Czapla, and Jeremy Howard. Applying a pre-trained language model to spanish twitter humor prediction. *arXiv preprint arXiv:1907.03187*, 2019.

[129] Jiaqi Liu, Ran Tong, Aowei Shen, Shuzheng Li, Changlin Yang, and Lisha Xu. Memeblip2: A novel lightweight multimodal system to detect harmful memes, 2025.

[130] Ankush Khandelwal. *Towards Identifying Humor and Author's Gender in Code-Mixed Social Media Content*. PhD thesis, Ph. D. Dissertation. International Institute of Information Technology Hyderabad, 2019.

[131] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. Ur-funny: A multimodal language dataset for understanding

humor. 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)*. Association for Computational Linguistics, 2019.

[132] Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh, Hunar Singh, and Vinay P. Namboodiri. Multimodal Humor Dataset: Predicting Laughter Tracks for Sitcoms. pages 576–585, 2021.

[133] Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-philippe Morency, and Soujanya Poria. M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 773–777, 2021.

[134] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. Amhuse: a multimodal dataset for humour sensing. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 438–445, 2017.

[135] Yuriel Ryan, Rui Yang Tan, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Humor in pixels: Benchmarking large multimodal models understanding of online comics. *arXiv preprint arXiv:2509.12248*, 2025.

[136] Nancy Bell and Salvatore Attardo. Failed humor: Issues in non-native speakers' appreciation and understanding of humor. *Intercultural Pragmatics*, 7(3), 2010.

[137] Eduardo Herrera-Alba and Rubén Manrique. Exploring multimodal humor detection in latin-american spanish with ls-funny. *SN Computer Science*, 6(7):770, 2025.

[138] Hongyu Guo, Wenbo Shang, Xueyao Zhang, Shubo Zhang, Xu Han, and Binyang Li. Much: A multimodal corpus construction for conversational humor recognition based on chinese sitcom. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11692–11698, 2024.

[139] Yongyi Kui. Applying pre-trained model and fine-tune to conduct humor analysis on spanish tweets. In *IberLEF@ SEPLN*, pages 844–851, 2021.

[140] Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. A crowd-annotated spanish corpus for humor analysis, 2018.

[141] Fateme Najafi-Lapavandani and Mohammad Shirali-Shahreza. Humor detection in persian: A transformers-based approach. *International Journal of Information and Communication Technology Research*, 15:56–62, 02 2023.

[142] Zefeng Li, Hongfei Lin, Liang Yang, Bo Xu, and Shaowu Zhang. Memeplate: A chinese multimodal dataset for humor understanding in meme templates. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 527–538. Springer, 2022.

[143] Dongyu Zhang, Heting Zhang, Xikai Liu, Hongfei Lin, and Feng Xia. Telling the whole story: A manually annotated Chinese dataset for the analysis of humor in jokes. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6402–6407, Hong Kong, China, November 2019. Association for Computational Linguistics.

[144] Zhenghan Yu, Xinyu Hu, and Xiaojun Wan. Cfunmodel: A" funny" language model capable of chinese humor generation and processing. *arXiv preprint arXiv:2503.20417*, 2025.

[145] Akshi Kumar, Abhishek Mallik, and Sanjay Kumar. Humourhindinet: Humour detection in hindi web series using word embedding and convolutional neural network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(7):1–21, 2024.

[146] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. Large dataset and language model fun-tuning for humor recognition. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy, July 2019. Association for Computational Linguistics.

[147] Sai Kartheek Reddy Kasu, Mohammad Zia Ur Rehman, Shahid Shafi Dar, Rishi Bharat Junghare, Dhanvin Sanjay Namboodiri, and Nagendra Kumar. D-humor: Dark humor understanding via multimodal open-ended reasoning – a benchmark dataset and method, 2025.

[148] Kazuki Kawamura, Kengo Nakai, and Jun Rekimoto. Manzaiset: A multimodal dataset of viewer responses to japanese manzai comedy, 2025.

[149] Hongyu Guo, Wenbo Shang, Xueyao Zhang, Shubo Zhang, Xu Han, and Binyang Li. MUCH: A multimodal corpus construction for conversational humor recognition based on Chinese sitcom. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of*

the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11692–11698, Torino, Italia, May 2024. ELRA and ICCL.

[150] Veedant Jain, Gabriel Kreiman, and Felipe dos Santos Alves Feitosa. Humordb: Can ai understand graphical humor?, 2025.

[151] Stanley Cao and Sonny Young. Predicting winning captions for weekly new yorker comics. *arXiv preprint arXiv:2407.18949*, 2024.

[152] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[153] Cheng Zhang and Hayato Yamana. Wuy at semeval-2020 task 7: Combining bert and naive bayes-svm for humor assessment in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1071–1076, 2020.

[154] Rutal Mahajan and Mukesh Zaveri. An automatic humor identification model with novel features from berger's typology and ensemble models. *Decision Analytics Journal*, 11:100450, 2024.

[155] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees, 2020.

[156] Vikramkumar, Vijaykumar B, and Trilochan. Bayes and naive bayes classifier, 2014.

[157] Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. Is This a Joke? Detecting Humor in Spanish Tweets. volume 10022, pages 139–150. 2016. arXiv:1703.09527 [cs].

[158] Rafal Rzepka, Yusuke Amaya, Motoki Yatsu, and Kenji Araki. Automatic narrative humor recognition method using machine learning and semantic similarity based punchline detection. In *International workshop on chance discovery, data synthesis and data market in IJCAI2015, IJCAI*, 2015.

[159] Arunima Jaiswal et al. Pun detection using soft computing techniques. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 5–9. IEEE, 2019.

[160] Thomas Winters and Pieter Delobelle. Dutch humor detection by generating negative examples. *arXiv preprint arXiv:2010.13652*, 2020.

[161] C Ramakristanaiah, P Namratha, Rajendra Kumar Ganiya, and Midde Ranjit Reddy. A survey on humor detection methods in communications. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 668–674. IEEE, 2021.

[162] Pariksha Prajapati, Arunima Jaiswal, Aastha, Shilpi, Neha, and Nitin Sachdeva. Empirical Analysis of Humor Detection Using Deep Learning and Machine Learning on Kaggle Corpus. In Vijayan Sugumaran, Divya Upadhyay, and Shanu Sharma, editors, *Advancements in Interdisciplinary Research*, pages 300–312, Cham, 2022. Springer Nature Switzerland.

[163] Marcio Lima Inácio, Gabriela Wick-Pedro, and Hugo Gonçalo Oliveira. What do humor classifiers learn? an attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 88–98, 2023.

[164] Hugo Gonçalo Oliveira, André Clemêncio, and Ana Alves. Corpora and baselines for humour recognition in Portuguese. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1278–1285, Marseille, France, May 2020. European Language Resources Association.

[165] Neamul Islam Fahim, Rifah Khan, Sujana Rahman, Nusrat Akter, and Mohammad Nurul Huda. Humor Detection Using Machine Learning Approach. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pages 1217–1222, Dhaka, Bangladesh, May 2024. IEEE.

[166] Nathan Chi and Ryan Chi. RedwoodNLP at SemEval-2021 Task 7: Ensembled Pretrained and Lightweight Models for Humor Detection. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1209–1214, Online, August 2021. Association for Computational Linguistics.

[167] Vijay Kumar, Ranjeet Walia, and Shivam Sharma. DeepHumor: a novel deep learning framework for humor detection. *Multimedia Tools and Applications*, 81(12):16797–16812, May 2022.

[168] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[169] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[170] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics Data Analysis*, 38(4):367–378, 2002. Nonlinear Methods and Data Mining.

[171] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[172] Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and applications*, 5(64-67):2, 2001.

[173] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[174] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[175] Dario Bertero and Pascale Fung. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, 2016.

[176] Jiahuan Yan, Yule Yang, and Xi Zhu. Humor prediction with bi-directional long-short term memory. In *2021 International Conference on Neural Networks, Information and Communication Engineering*, volume 11933, pages 110–116. SPIE, 2021.

[177] Ashish Kayastha and Alexander Redei. That's what she said: Humor identification with word embeddings and recurrent neural networks. In *Future of Information and Communication Conference*, pages 209–221. Springer, 2022.

[178] Rida Miraj and Masaki Aono. Combining bert and multiple embedding methods with the deep neural network for humor detection. In *International Conference on Knowledge Science, Engineering and Management*, pages 53–61. Springer, 2021.

[179] Dario Bertero and Pascale Fung. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 496–501, 2016.

[180] Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[181] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[182] Lei Chen and Chong Min Lee. Convolutional neural network for humor recognition. *arXiv preprint arXiv:1702.02584*, 2017.

[183] Peng-Yu Chen and Von-Wun Soo. Humor recognition using deep learning. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)*, pages 113–117, 2018.

[184] Lei Chen and Chong MIn Lee. Predicting audience's laughter using convolutional neural network. *arXiv preprint arXiv:1702.02584*, 2017.

[185] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, June 2017. arXiv:1706.03762 [cs] version: 1.

[186] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025.

[187] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[188] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[189] Andrew Pulver and Siwei Lyu. Lstm with working memory. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 845–851. IEEE, 2017.

[190] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[191] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[192] Artem V Slobodianiuk and Serhiy O Semerikov. Advances in neural text generation: A systematic review. 2025.

[193] Tyler Loakman, William Thorne, and Chenghua Lin. Who's laughing now? an overview of computational humour generation and explanation, 2025.

[194] Tyler Loakman, William Thorne, and Chenghua Lin. Comparing apples to oranges: A dataset & analysis of llm humour understanding from traditional puns to topical jokes. *arXiv preprint arXiv:2507.13335*, 2025.

[195] Shreyas Joshi, Muhammad Shahnawaz Khan, Aditya Dafe, Kavita Singh, Vedant Zope, and Tanish Jhamtani. Fine tuning llms for low resource languages. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 511–519. IEEE, 2024.

[196] Jisu An, Junseok Lee, Jeoungeun Lee, and Yongseok Son. Towards llm-centric multimodal fusion: A survey on integration strategies and techniques, 2025.

[197] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[198] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[199] Kenneth Ward Church, Zeyu Chen, and Yanjun Ma. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27(6):763–778, 2021.

[200] Marcio Lima Inácio and Hugo Gonçalo Oliveira. Exploring multimodal models for humor recognition in portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 568–574, 2024.

[201] Aishwarya Gupta, Avik Pal, Bholeshwar Khurana, Lakshay Tyagi, and Ashutosh Modi. Humor@ iitk at semeval-2021 task 7: Large language models for quantifying humor and offensiveness. *arXiv preprint arXiv:2104.00933*, 2021.

[202] Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. Can pre-trained language models understand chinese humor? In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 465–480, 2023.

[203] Shih-Hung Wu, Tsz-Yeung Lau, and Yu-Feng Huang. Humour classification according to genre and technique by fine-tuning llms. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 156–169. Springer, 2025.

[204] Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17826–17834, 2024.

[205] Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny large language models. *arXiv preprint arXiv:2403.00794*, 2024.

[206] Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy T Rogers, Kevin G Jamieson, et al. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *Advances in Neural Information Processing Systems*, 37:125264–125286, 2024.

[207] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023.

[208] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[209] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[210] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155, 2023.

[211] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[212] Petra Bago and Nikola Bakarić. Few-shot prompting, full-scale confusion: Evaluating large language models for humor detection in croatian tweets. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 9–16, 2025.

[213] Xu Guo, Han Yu, Boyang Li, Hao Wang, Pengwei Xing, Siwei Feng, Zaiqing Nie, and Chunyan Miao. Federated learning for personalized humor recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–18, 2022.

[214] Liana Ermakova, Ricardo Campos, Anne-Gwenn Bosser, and Tristan Miller. Overview of JOKER: Humour in the machine. In Jorge Carrillo de Albornoz, Julio Gonzalo, Laura Plaza, Alba García Seco de Herrera, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings*, volume 16089 of *Lecture Notes in Computer Science*, pages 315–337, Cham, 2025. Springer.

[215] Zhongren Dong, Donghao Wang, Ciqiang Chen, Dong-Yan Huang, and Zixing Zhang. Mhsdb: A comprehensive benchmark for multimodal humor and sarcasm detection leveraging foundation models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[216] Orion Weller and Kevin Seppi. The rjokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, 2020.

[217] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4027–4032, 2019.