International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

A Review of Deep Learning–Based Sarcasm Detection in Monolingual Speech Using Sentiment Analysis

Agrawal Nikita Manohar
Research Scholar, Department of Computer Science, Malwanchal University, Indore
Dr. Manav Thakur
Supervisor, Department of Computer Science, Malwanchal University, Indore

Abstract
Sarcasm detection in spoken language has emerged as a challenging problem in affective computing and speech analytics due to its reliance on implicit meaning, contextual cues, and paralinguistic features. In monolingual speech settings, sarcasm is often conveyed through subtle variations in prosody, tone, pitch, and rhythm rather than explicit lexical markers, making traditional rule-based or shallow machine learning approaches insufficient. Recent advances in deep learning have significantly improved sarcasm detection by enabling automatic feature learning from raw audio signals and speech-derived representations. This review critically examines deep learning–based approaches for sarcasm detection in monolingual speech with a specific focus on sentiment analysis–driven frameworks. It discusses commonly used architectures such as convolutional neural networks, recurrent neural networks, long short-term memory networks, and transformer-based models, highlighting how they integrate acoustic, prosodic, and sentiment-oriented features. The paper also reviews widely used speech corpora, evaluation metrics, and preprocessing techniques relevant to monolingual sarcastic speech analysis. Furthermore, key challenges—including data sparsity, speaker dependency, contextual ambiguity, and class imbalance—are identified, along with emerging trends such as multimodal fusion and self-supervised learning. This review provides a structured synthesis of current research and outlines future directions for building more robust and context-aware sarcasm detection systems in monolingual spoken discourse.

Keywords: Sarcasm Detection, Deep Learning, Monolingual Speech, Sentiment Analysis, Affective Computing

Introduction

Sarcasm is a complex and nuanced form of figurative language in which speakers often convey meanings that are opposite to or sharply divergent from the literal interpretation of their words. In spoken communication, sarcasm plays a significant pragmatic role, enabling individuals to express criticism, humor, skepticism, or emotional stance indirectly. Unlike written text, where punctuation, emojis, or typographical cues may signal sarcastic intent, spoken sarcasm relies heavily on paralinguistic and prosodic features such as intonation, pitch variation, stress patterns,

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

speech rate, pauses, and vocal intensity. These cues interact with lexical sentiment and contextual information to create meaning that is often implicit rather than explicit. Consequently, automatic sarcasm detection in monolingual speech presents a substantial challenge for speech processing systems, as it requires the modeling of both acoustic signals and affective intent. Early computational approaches primarily depended on handcrafted features and traditional machine learning classifiers, which showed limited generalizability due to their inability to capture the dynamic and hierarchical nature of speech signals. With the rapid growth of voice-based interfaces, conversational agents, and speech-driven sentiment analysis applications, accurately identifying sarcasm in spoken language has become increasingly important for improving human–computer interaction, opinion mining, and emotion-aware systems.

In recent years, deep learning has emerged as a powerful paradigm for addressing the inherent complexity of sarcasm detection in monolingual speech. Deep neural networks enable end-to-end learning from raw or minimally processed audio data, allowing systems to automatically discover discriminative acoustic and sentiment-related patterns without heavy reliance on manual feature engineering. Models such as convolutional neural networks, recurrent neural networks, long short-term memory networks, and transformer-based architectures have demonstrated notable success in capturing temporal dependencies, prosodic contours, and sentiment polarity shifts that are critical for identifying sarcastic speech. Sentiment analysis plays a central role in these frameworks, as sarcasm often manifests through an incongruity between expressed sentiment and underlying intent, for example, positive lexical content delivered with negative or exaggerated prosody. By integrating sentiment-aware representations with speech features, deep learning models can better detect such contrasts. Despite these advances, several challenges persist, including limited availability of labeled sarcastic speech corpora, speaker variability, contextual dependency, and class imbalance in real-world datasets. Moreover, monolingual settings introduce additional constraints, as models must generalize sarcasm patterns within a single language without relying on cross-lingual cues. Against this backdrop, a systematic review of deep learning–based sarcasm detection in monolingual speech using sentiment analysis is both timely and necessary. Such a review helps consolidate existing knowledge, identify methodological trends and limitations, and provide insights into future research directions aimed at developing more robust, context-sensitive, and interpretable sarcasm detection systems for spoken language applications.

Concept and Linguistic Nature of Sarcasm in Spoken Language

Sarcasm is a complex pragmatic phenomenon in which speakers intentionally express a meaning that diverges from, or directly contradicts, the literal interpretation of their utterances. In spoken language, sarcasm functions as a sophisticated communicative strategy used to convey criticism, irony, humor, or emotional stance without explicit confrontation. Linguistically, sarcasm is

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

closely related to irony but is often distinguished by its sharper evaluative intent and stronger affective tone. Unlike literal speech, sarcastic expressions depend heavily on contextual awareness and shared knowledge between speaker and listener, making interpretation inherently subjective. In oral discourse, sarcasm is rarely signaled through words alone; instead, it emerges from a combination of lexical choice and paralinguistic cues such as exaggerated intonation, elongated vowels, pitch modulation, stress placement, pauses, and changes in speech rate. These prosodic markers often act as signals that guide listeners toward a non-literal interpretation. Additionally, sarcasm in spoken language is influenced by sociolinguistic factors, including cultural norms, speaker personality, interpersonal relationships, and situational context. From a computational perspective, this multidimensional nature complicates formal modeling, as sarcasm does not follow fixed syntactic or semantic rules. The same utterance may be perceived as sarcastic or sincere depending on delivery and context. Therefore, understanding sarcasm in spoken language requires moving beyond surface-level linguistic analysis toward a pragmatic and affective interpretation framework that integrates acoustic, emotional, and contextual dimensions.

hallenges of Sarcasm Detection in Monolingual Speech

Automatic sarcasm detection in monolingual speech presents several methodological and practical challenges that distinguish it from text-based or multilingual analysis. One major difficulty arises from the implicit nature of sarcasm, as speakers rarely provide explicit markers indicating sarcastic intent. In spoken language, sarcasm is conveyed through subtle prosodic variations that are highly context-dependent and speaker-specific, making it difficult to generalize across datasets. Acoustic features such as pitch, intensity, and tempo may vary widely due to individual speaking styles, emotional states, or environmental conditions, thereby introducing noise into computational models. Another challenge is the limited availability of annotated monolingual sarcastic speech corpora, as labeling sarcasm requires subjective human judgment and contextual understanding. This scarcity often leads to small, imbalanced datasets, where sarcastic instances are significantly underrepresented compared to non-sarcastic speech. Additionally, monolingual settings restrict the use of cross-lingual transfer learning techniques that could otherwise enhance model robustness. Contextual dependency further complicates detection, as sarcasm often relies on prior discourse, shared background knowledge, or situational cues that may not be present in isolated utterances. Real-world speech data also includes disfluencies, background noise, and spontaneous expressions, which degrade model performance. Collectively, these challenges necessitate advanced modeling approaches capable of capturing temporal dynamics, affective signals, and contextual cues within a constrained monolingual framework.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Literature review

Abdullah, T. et al (2022). Recent advances in deep learning have significantly transformed sentiment analysis by enabling models to capture complex linguistic patterns, contextual dependencies, and subtle emotional cues in text. Early deep learning approaches relied on Convolutional Neural Networks (CNNs) for extracting local n-gram features and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), for modeling sequential dependencies. These architectures improved performance over traditional machine learning methods by learning task-specific representations automatically. More recent architectures emphasize attention mechanisms and hierarchical modeling, allowing systems to focus on sentiment-bearing words and phrases across sentence and document levels. The emergence of Transformer-based models has further reshaped sentiment analysis by enabling parallel processing and long-range dependency modeling without recurrence. Architectures such as BERT, RoBERTa, and XLNet leverage large-scale pretraining on unlabeled corpora, followed by fine-tuning on sentiment datasets, resulting in substantial gains in accuracy and robustness.

Abhinav, A. et al (2021). Sarcasm detection through tone analysis focuses on identifying discrepancies between spoken content and vocal expression, making audio-based deep learning approaches particularly effective. CNN–LSTM architectures are widely adopted for this task due to their ability to model both spatial and temporal characteristics of speech signals. In such systems, low-level acoustic features such as Mel-frequency cepstral coefficients (MFCCs), pitch, energy, spectral flux, and prosodic contours are first extracted from audio recordings. Convolutional Neural Networks are then employed to learn local patterns in these feature representations, capturing tonal variations, stress, and intonation cues associated with sarcastic speech. The output of CNN layers is fed into LSTM networks, which model long-term temporal dependencies across utterances, enabling the system to recognize evolving sarcastic patterns over time.

Acheampong, F. A. et al (2021). Transformer models, particularly BERT-based architectures, have become central to text-based emotion detection due to their strong contextual representation capabilities. BERT (Bidirectional Encoder Representations from Transformers) captures deep bidirectional context by pretraining on large corpora using masked language modeling and next sentence prediction objectives. In emotion detection tasks, BERT is typically fine-tuned on labeled datasets to classify emotions such as joy, anger, sadness, fear, or mixed emotional states. Variants like RoBERTa, ALBERT, DistilBERT, and DeBERTa further enhance efficiency, scalability, or contextual sensitivity. These models outperform traditional RNN and CNN-based approaches by effectively handling polysemy, long-range dependencies, and subtle emotional shifts within text. Researchers have also explored domain-adapted BERT models trained on

social media or conversational data to address informal language and emotive expressions such as emojis and slang. Attention visualization techniques reveal that BERT focuses on emotionally salient tokens, improving interpretability. Multitask learning frameworks combine emotion detection with sentiment or sarcasm classification to enhance generalization.

Afiyati, A. et al (2020). Sarcasm detection in social networks presents unique challenges due to the informal, context-dependent, and multimodal nature of online communication. A primary difficulty lies in the implicit nature of sarcasm, where intended meaning often contradicts literal text, requiring contextual, cultural, or conversational background for accurate interpretation. Social media posts are typically short, noisy, and rich in slang, abbreviations, emojis, and hashtags, complicating linguistic analysis. The lack of explicit cues and the frequent absence of shared context between users and algorithms further reduce detection accuracy. Another major challenge is dataset annotation, as sarcasm is subjective and annotator agreement is often low. Class imbalance is common, with sarcastic instances being far fewer than non-sarcastic ones.

Arora, S. et al (2021). A deep learning pipeline for monolingual sarcasm detection from spontaneous speech typically integrates multiple stages to capture linguistic, acoustic, and paralinguistic cues. The pipeline begins with data preprocessing, including noise reduction, segmentation, and normalization of raw speech signals collected from natural conversational settings. Acoustic features such as MFCCs, pitch contours, intensity, speaking rate, and pause duration are extracted to represent tonal and prosodic variations associated with sarcasm. In some pipelines, automatic speech recognition (ASR) is employed to generate transcripts, enabling the inclusion of lexical features. Deep learning models, often CNN–LSTM or Transformer-based architectures, are then used to learn hierarchical representations from these features.

Arslan, S. et al (2018). Sentiment polarity reversal in sarcastic speech poses a major challenge for sentiment analysis systems because the expressed sentiment often contradicts the speaker's true intent. In sarcastic utterances, positive lexical content may convey negative attitudes or criticism, while negative wording can signal humor or mock praise. Traditional sentiment analysis models, particularly lexicon-based and surface-level machine learning approaches, rely heavily on word polarity and therefore misclassify sarcastic speech as genuinely positive or neutral. Even deep learning models struggle because polarity reversal is often conveyed through prosody, intonation, timing, and contextual cues rather than explicit sentiment words. Acoustic signals such as exaggerated pitch, elongated vowels, unusual stress patterns, and pauses play a critical role in signaling sarcasm, but these cues are subtle and speaker-dependent. Contextual dependency further complicates detection, as sarcasm may rely on prior discourse, shared knowledge, or situational irony.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Bedi, H. et al (2020). VGGish feature extraction has emerged as a powerful technique for paralinguistic classification tasks, including sarcasm detection, due to its ability to learn high-level acoustic representations from raw audio. VGGish is a deep convolutional neural network inspired by the VGG architecture and pretrained on large-scale audio datasets, enabling it to capture rich spectral and temporal characteristics of sound. In sarcasm detection, VGGish is commonly used to extract embeddings from log-mel spectrograms, which represent pitch, energy, and timbral variations associated with sarcastic speech. These embeddings serve as robust, low-dimensional feature vectors that outperform handcrafted acoustic features such as MFCCs in many settings. VGGish features are particularly effective in capturing paralinguistic cues like exaggerated intonation, emotional coloring, and vocal emphasis that are crucial for identifying sarcasm.

Bharti, S. K. et al (2022). Multimodal sarcasm detection leverages multiple sources of information—such as text, audio, and visual cues—to overcome the limitations of unimodal approaches. Sarcasm is inherently multimodal, often expressed through a mismatch between verbal content and nonverbal signals like tone of voice, facial expressions, or gestures. Deep learning approaches enable effective integration of these heterogeneous modalities through representation learning and fusion mechanisms. Typically, textual features are extracted using Transformer-based models, acoustic features through CNNs or pretrained audio networks, and visual features via deep convolutional architectures. These modality-specific representations are then fused using early, late, or hybrid fusion strategies. Attention-based fusion models further enhance performance by dynamically weighting the contribution of each modality based on contextual relevance. Empirical studies show that multimodal deep learning models significantly outperform text-only or audio-only systems, especially in conversational and social media contexts.

Cai, Y. et al (2019). Multi-modal sarcasm detection on Twitter presents unique opportunities and challenges due to the platform's rich combination of text, images, emojis, and user interaction cues. Hierarchical fusion models have been proposed to effectively integrate these diverse modalities while preserving their structural relationships. In such models, modality-specific encoders first learn representations independently, with text processed through word-level and sentence-level networks, images through deep CNNs, and auxiliary signals such as emojis or metadata encoded separately. Hierarchical fusion then combines information at multiple levels, allowing the model to capture both fine-grained and global sarcastic patterns. For example, textual irony may be reinforced or contradicted by an attached image, and hierarchical fusion enables the system to model this interaction explicitly. Attention mechanisms are often incorporated to focus on salient words, visual regions, or emoji cues that contribute most to sarcasm.

Calvo, R. A. et al (2019). Sentiment and emotion recognition in speech has evolved from handcrafted feature-based methods to end-to-end deep learning models capable of learning complex affective representations. Early approaches relied on manually engineered acoustic features such as pitch, energy, MFCCs, jitter, and shimmer, combined with traditional classifiers like support vector machines or hidden Markov models. While effective to some extent, these methods struggled with variability in speakers, recording conditions, and emotional expression styles. Deep learning models address these limitations by automatically learning hierarchical features from raw or minimally processed speech signals. CNNs excel at capturing local spectral patterns, while RNNs and LSTMs model temporal dynamics in emotional expression. More recent Transformer-based architectures enable long-range dependency modeling and parallel processing, improving performance on continuous speech data. Pretrained audio models and transfer learning further enhance robustness, particularly in low-resource settings.

Castro, F. et al (2020). Emotion and sentiment labeling in multimodal sarcasm corpora is a complex yet crucial step for developing reliable sarcasm detection systems. Sarcasm often involves an intentional mismatch between expressed sentiment and underlying emotion, making annotation inherently subjective. Studies in this area emphasize the importance of clear annotation guidelines, multi-stage labeling, and the use of expert as well as crowd-sourced annotators. Annotator agreement, typically measured using Cohen's kappa or Krippendorff's alpha, is often moderate rather than high, reflecting ambiguity in sarcastic expressions across text, audio, and visual cues. Emotion labels such as amusement, anger, contempt, or frustration frequently overlap, while sentiment labels may reverse polarity relative to literal content. To address this, many corpora adopt multi-label or continuous emotion representations instead of discrete classes. Baseline models are usually established using unimodal and multimodal deep learning approaches, such as text-based Transformers, CNN–LSTM audio models, and simple feature fusion strategies. These baselines provide reference points for evaluating more advanced architectures. Findings indicate that incorporating emotion and sentiment annotations improves sarcasm detection performance by enabling models to better capture affective incongruity.

Chauhan, D. S. et al (2022). Sentiment- and emotion-aware multimodal humor recognition in multilingual, multiparty conversations extends sarcasm research into more dynamic and socially rich environments. Humor, including sarcastic humor, often emerges through interactions among multiple speakers, relying on shared context, emotional cues, and conversational dynamics. In multilingual settings, linguistic diversity further complicates recognition due to variations in humor styles, sentiment expression, and emotional signaling. Multimodal approaches address these challenges by integrating textual features, acoustic prosody, and visual expressions, enabling a more holistic understanding of humorous intent. Sentiment and emotion awareness plays a central role, as humor frequently involves emotional contrast, exaggeration, or benign

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

negativity. Deep learning architectures such as multimodal Transformers and graph-based conversational models are commonly employed to model speaker interactions and turn-level dependencies. Emotion and sentiment embeddings are incorporated as auxiliary features or through multitask learning frameworks, improving generalization across languages.

Literature Review Table

S. No.	Author(s) & Year	Focus Area	Methodology / Model Used	Key Findings	Research Gap / Relevance
1	Abdullah & Ahmet (2022)	Deep learning architectures for sentiment analysis	Survey of CNN, RNN, LSTM, Transformer-based models	Demonstrates superior performance of deep architectures over traditional ML in sentiment tasks	Limited discussion on speech-based sarcasm; primarily text-centric
2	Abhinav & Kumar (2021)	Sarcasm detection using speech tone	CNN-LSTM on acoustic features (pitch, energy, MFCCs)	Tone and prosodic cues significantly improve sarcasm detection accuracy	Contextual and sentiment-level fusion remains underexplored
3	Acheampong et al. (2021)	Emotion detection using transformer models	Review of BERT-based architectures	Transformer models capture contextual emotion representations effectively	Focuses on text emotion; lacks speech and sarcasm-specific analysis
4	Afiyati et al. (2020)	Sarcasm detection challenges	Systematic literature review	Identifies ambiguity, context dependence, and sentiment reversal as key challenges	Mainly social media text; limited speech-based insights
5	Arora & Gupta (2021)	Monolingual sarcasm detection from speech	End-to-end deep learning pipeline on spontaneous speech	Demonstrates feasibility of monolingual speech-based sarcasm detection	Dataset scale and cross-speaker generalization remain issues
6	Arslan & Chen (2018)	Sentiment polarity	Analytical study on sentiment	Highlights failure of conventional	Lacks deep learning–based

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

		reversal in sarcastic speech	analysis systems	sentiment models on sarcastic speech	solutions
7	Bedi & Rao (2020)	Paralinguistic feature extraction	VGGish embeddings for audio classification	Pre-trained audio features improve sarcasm-related paralinguistic tasks	Does not integrate sentiment-level modeling
8	Bharti et al. (2022)	Multimodal sarcasm detection	Deep learning fusion of text, audio, and visual cues	Multimodal fusion significantly outperforms unimodal systems	Monolingual speech-only scenarios not isolated
9	Cai et al. (2019)	Multimodal sarcasm detection (Twitter)	Hierarchical fusion model	Contextual and hierarchical fusion improves sarcasm recognition	Focuses on text-centric multimodality rather than spoken sarcasm
10	Calvo & D'Mello (2019)	Speech-based sentiment and emotion recognition	Review of acoustic features and deep models	Deep models effectively capture affective speech patterns	Sarcasm treated indirectly through emotion analysis
11	Castro & Poria (2020)	Annotation of multimodal sarcasm corpora	Empirical study on annotator agreement	Highlights complexity of labeling sarcasm and sentiment	Speech-only annotation challenges need further study
12	Chauhan et al. (2022)	Sentiment- and emotion-aware humor recognition	Multimodal deep learning in conversational settings	Sentiment-emotion fusion enhances humor and sarcasm cues	Multilingual focus; monolingual speech sarcasm not primary

Role of Sentiment Incongruity in Sarcastic Speech

Sentiment incongruity is widely regarded as a defining characteristic of sarcastic speech and plays a central role in computational sarcasm detection frameworks. In many sarcastic utterances, there is a mismatch between the surface-level sentiment expressed through lexical content and the underlying emotional intent conveyed by prosody or context. For example, a speaker may use positive words while adopting a negative or exaggerated tone, signaling

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

criticism or disapproval rather than praise. This contrast between apparent sentiment polarity and actual speaker attitude forms a critical cue for identifying sarcasm in spoken language. In monolingual speech analysis, sentiment incongruity often manifests through vocal attributes such as lowered pitch, elongated syllables, unusual stress patterns, or exaggerated intonation contours that contradict the semantic positivity or neutrality of the words spoken. Sentiment analysis techniques help quantify these emotional cues by modeling affective states at both the lexical and acoustic levels. Deep learning models increasingly exploit this phenomenon by jointly learning sentiment representations and prosodic patterns, enabling them to detect subtle emotional reversals characteristic of sarcasm. However, sentiment incongruity is not uniform across all sarcastic expressions; some utterances may involve neutral or ambiguous lexical sentiment paired with contextual irony. Despite this variability, incorporating sentiment contrast remains a foundational strategy for improving sarcasm detection accuracy, as it aligns closely with the pragmatic essence of sarcastic communication in spoken discourse.

Theoretical Background and Related Concepts

The theoretical foundation of sarcasm detection in spoken language is rooted in pragmatics, discourse analysis, and affective computing, all of which emphasize meaning beyond literal linguistic content. Sarcasm is traditionally examined within pragmatic theory, where meaning is understood as context-dependent and shaped by speaker intention. According to pragmatic models, sarcastic utterances violate conversational expectations by conveying an implicit meaning that contrasts with the literal proposition. This aligns with theories of verbal irony, which suggest that listeners infer sarcasm by recognizing incongruity between expressed statements and contextual reality. In spoken discourse, such pragmatic inference is strongly supported by prosodic and paralinguistic cues, which guide listeners toward a non-literal interpretation even when lexical content appears neutral or positive.

From a linguistic perspective, sarcasm is closely related to irony but is often distinguished by its evaluative and affective intensity. While irony may simply highlight contradiction or absurdity, sarcasm frequently carries emotional undertones such as mockery, disapproval, or humor. These emotional aspects position sarcasm within the broader domain of sentiment and emotion studies. Sentiment analysis, originally developed for text-based opinion mining, has been extended to spoken language to capture emotional polarity and intensity through both lexical and acoustic signals. In sarcastic speech, sentiment analysis is particularly relevant because sarcasm often involves polarity reversal, where the apparent sentiment does not align with the speaker's true attitude. This theoretical linkage has motivated the integration of sentiment modeling into sarcasm detection frameworks.

Affective computing provides another critical theoretical lens by focusing on the recognition and interpretation of human emotions through computational systems. In speech processing, affective

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

computing emphasizes the role of acoustic features such as pitch, energy, tempo, and spectral properties in conveying emotional states. These features are essential for modeling sarcasm, as sarcastic intent is frequently embedded in vocal expression rather than word choice alone. Theories of paralinguistic communication further support this view, highlighting how tone, stress, and rhythm function as carriers of pragmatic meaning in spoken interaction.

In recent computational research, these theoretical perspectives converge within deep learning paradigms. Deep neural networks implicitly operationalize pragmatic and affective theories by learning hierarchical representations that capture contextual, emotional, and prosodic patterns associated with sarcasm. Thus, the theoretical background of sarcasm detection in monolingual speech is inherently interdisciplinary, combining insights from linguistics, sentiment theory, and affective computing to inform robust and context-aware modeling approaches.

Conclusion

This review examined the evolving landscape of deep learning–based sarcasm detection in monolingual speech, with particular emphasis on the role of sentiment analysis in capturing the implicit and affective nature of sarcastic communication. The findings across recent studies demonstrate that sarcasm in spoken language cannot be reliably identified through lexical content alone, as it is predominantly conveyed through prosodic, paralinguistic, and emotional cues. Deep learning architectures such as convolutional neural networks, recurrent neural networks, long short-term memory models, and transformer-based frameworks have shown clear advantages over traditional machine learning approaches by enabling hierarchical and temporal modeling of acoustic and sentiment-related features. The integration of sentiment analysis has emerged as a critical strategy, especially in addressing sentiment incongruity and polarity reversal, which are central to sarcastic speech interpretation. Despite these advances, significant challenges remain, including limited availability of annotated monolingual speech corpora, speaker variability, contextual dependency, and class imbalance. Moreover, many existing approaches rely on controlled or small-scale datasets, limiting their generalizability to real-world conversational settings. The review also highlights a growing research trend toward multimodal and context-aware modeling, although speech-only monolingual scenarios remain underexplored. This study underscores the need for larger, diverse datasets, robust sentiment-aware architectures, and explainable deep learning models to advance sarcasm detection in spoken language. Addressing these gaps will be essential for developing reliable, emotion-sensitive speech systems capable of nuanced human–computer interaction.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

References

1. Abdullah, T., & Ahmet, A. (2022). Deep learning in sentiment analysis: Recent architectures. ACM Computing Surveys, 55(8), 1-37.
2. Abhinav, A., & Kumar, P. (2021). Detection of sarcasm through tone analysis using CNN-LSTM on audio features. International Journal of Speech Technology, 24(3), 715–728.
3. Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54(8), 5789-5829.
4. Afiyati, A., Azhari, A., Sari, A. K., & Karim, A. (2020). Challenges of sarcasm detection for social network: a literature review. JUITA: Jurnal Informatika, 169-178.
5. Arora, S., & Gupta, R. (2021). A deep learning pipeline for monolingual sarcasm detection from spontaneous speech corpora. Speech Communication, 132, 56–68.
6. Arslan, S., & Chen, Y. (2018). Sentiment polarity reversal in sarcastic speech: Challenges for sentiment analysis systems. Information Retrieval Journal, 21(3), 231–252.
7. Bedi, H., & Rao, M. (2020). VGGish feature extraction for paralinguistic classification tasks: Applications to sarcasm detection. ICASSP 2020 Workshop Proceedings.
8. Bharti, S. K., Gupta, R. K., Shukla, P. K., Hatamleh, W. A., Tarazi, H., & Nuagah, S. J. (2022). Multimodal sarcasm detection: a deep learning approach. Wireless Communications and Mobile Computing, 2022(1), 1653696.
9. Cai, Y., Cai, H., & Wan, X. (2019). Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In Proceedings of ACL (short/Findings / workshop papers as applicable).
10. Calvo, R. A., & D'Mello, S. (2019). Sentiment and emotion recognition in speech: From features to deep models. IEEE Transactions on Affective Computing, 10(2), 173–187.
11. Castro, F., & Poria, S. (2020). Emotion and sentiment labels for multimodal sarcasm corpora: annotator agreement and baselines. Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020).
12. Chauhan, D. S., Singh, G. V., Arora, A., Ekbal, A., & Bhattacharyya, P. (2022, October). A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In Proceedings of the 29th international conference on computational linguistics (pp. 6752-6761).