International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Sarcasm Detection in Monolingual Speech Using Deep Learning–Driven 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 is a challenging task due to the implicit and often contradictory relationship between literal expression and intended sentiment. In audio-based communication, sarcasm is primarily conveyed through acoustic and prosodic cues such as pitch modulation, intonation patterns, speech rate, and energy variation rather than explicit lexical indicators. This study presents a deep learning–based approach for sarcasm detection using sentiment analysis of a monolingual audio corpus. The proposed framework focuses on extracting sentiment-aware acoustic features, including Mel-frequency cepstral coefficients, prosodic features, and spectral characteristics, which are then modeled using deep neural architectures to capture both spatial and temporal dependencies in speech signals. By exploiting sentiment incongruity between vocal expression and underlying intent, the system aims to distinguish sarcastic utterances from non-sarcastic ones more effectively. Experimental evaluation demonstrates that integrating sentiment-oriented features with deep learning models significantly enhances sarcasm detection performance, highlighting the potential of audio-based sentiment analysis for improving speech-driven intelligent systems.

Keywords: Sarcasm Detection, Sentiment Analysis, Audio Corpus, Deep Learning, Speech Processing

Introduction

Sarcasm is a complex and nuanced form of expression in which the speaker’s intended meaning differs sharply from the literal interpretation of the spoken words, often conveying ridicule, irony, or mock praise. In spoken communication, sarcasm is not solely dependent on lexical content but is strongly influenced by vocal cues such as intonation, pitch variation, stress, tempo, and energy patterns. This makes sarcasm detection particularly challenging for automated systems, especially when compared to text-based analysis. With the rapid growth of voice-driven technologies—such as virtual assistants, conversational agents, call center analytics, and emotion-aware human–computer interaction systems—the ability to accurately detect sarcasm in speech has become increasingly important. Traditional sentiment analysis methods often fail in sarcastic scenarios because sarcastic speech frequently exhibits a mismatch between expressed sentiment and intended sentiment, for example, positive words delivered with negative or exaggerated prosody. This sentiment incongruity is a key signal that can be exploited for sarcasm detection. Recent advances in deep learning have significantly improved performance in speech and affective computing tasks by enabling models to learn hierarchical and discriminative representations directly from raw or low-level acoustic features. Deep neural architectures such as convolutional neural networks and recurrent neural networks are particularly effective in modeling spectral patterns and temporal dependencies present in audio signals. In a monolingual setting, where linguistic variability is constrained to a single language, it becomes feasible to focus more deeply on acoustic-prosodic cues and sentiment-related patterns without the added complexity of cross-lingual variation. This study explores sarcasm detection based on sentiment analysis of a monolingual audio corpus using deep learning techniques, aiming to capture the interaction between vocal sentiment expression and sarcastic intent. By integrating sentiment-aware acoustic features with robust deep learning models, the proposed approach seeks to improve the automatic identification of sarcasm in spoken language, thereby contributing to more context-aware, emotionally intelligent, and socially sensitive speech-based artificial intelligence systems.

Scope of the Study

The scope of this study is confined to the detection of sarcasm in spoken language through sentiment analysis of a monolingual audio corpus using deep learning techniques. The research focuses exclusively on audio-based features, emphasizing acoustic and prosodic characteristics such as pitch, intensity, intonation, speech rate, and spectral properties, while excluding textual, visual, or multimodal inputs. The study is limited to a single language in order to reduce linguistic variability and to enable a more precise analysis of sentiment incongruity as expressed through vocal cues. Deep learning models are employed to learn discriminative patterns associated with sarcastic and non-sarcastic speech, with particular attention given to sentiment-oriented feature representations. The work aims to evaluate model performance under controlled experimental conditions and does not attempt real-time deployment or cross-domain generalization. Consequently, the findings are primarily applicable to research contexts, speech analytics, and emotion-aware systems operating within similar monolingual and audio-centric settings.

Challenges in Audio-Based Sarcasm Detection

Audio-based sarcasm detection presents several inherent challenges due to the subtle and subjective nature of sarcastic expression in speech. One major difficulty lies in the high variability of acoustic cues such as pitch, intonation, and speech rate, which can differ significantly across speakers, emotional states, and speaking styles. Sarcasm often manifests through nuanced prosodic patterns rather than consistent or universal markers, making feature generalization complex. Additionally, the overlap between sarcastic speech and other expressive states such as humor, frustration, or boredom can lead to ambiguity and misclassification. The absence of contextual and lexical information further complicates detection, as sentiment incongruity must be inferred solely from vocal delivery. Limited availability of well-annotated sarcastic audio corpora and the subjective nature of labeling sarcasm introduce data imbalance and annotation bias. Moreover, background noise, recording quality, and speaker-dependent characteristics can adversely affect feature extraction and model robustness, posing significant challenges for reliable audio-based sarcasm recognition systems.

Literature review

Sarcasm detection has emerged as a critical subfield within sentiment analysis and affective computing due to sarcasm's inherent capacity to invert literal sentiment and undermine surface-level polarity cues. Early computational approaches struggled with sarcasm because sarcastic expressions often encode a contrast between explicit lexical sentiment and implicit speaker intent. Ghosh and Veale (2018) provide one of the most influential theoretical framings by conceptualizing sarcasm as a *sentiment shift phenomenon*, where positive lexical cues mask an underlying negative intent or vice versa. This framing proved foundational, as it enabled sarcasm detection to be treated not merely as a classification problem but as a pragmatic and discourse-level challenge. Joshi, Fersini, and Rosso (2018), in their comprehensive survey, further systematize sarcasm research by categorizing approaches into rule-based, machine learning, and deep learning paradigms, while emphasizing the linguistic properties of sarcasm such as incongruity, hyperbole, and contextual dependency. These studies collectively underscore that sarcasm is not an isolated textual feature but a pragmatic act embedded within broader communicative contexts, necessitating models that move beyond bag-of-words representations toward discourse-aware and cognitively informed frameworks.

Initial sarcasm detection systems predominantly relied on surface-level textual features, including sentiment lexicons, punctuation patterns, and emotive markers. However, such models demonstrated limited generalizability across domains and genres. Hazarika et al. (2018) marked a significant shift by introducing CASCADE, a context-aware deep learning framework that incorporated conversational history, user embeddings, and discourse structure in online discussion forums. Their work empirically demonstrated that sarcasm interpretation is highly dependent on prior utterances and speaker behavior. Similarly, Das and Kolya (2021) extended sarcasm detection into literary and pop-culture texts, revealing that humor literature exhibits sarcasm patterns distinct from social media discourse. These findings challenged the assumption of domain universality and highlighted the importance of genre-sensitive modeling. Moreover, FigLang 2020 provided standardized benchmarks and shared tasks, catalyzing methodological rigor and comparative evaluation across models. Collectively, these studies illustrate the field's transition from isolated sentence-level analysis toward richer contextual modeling, acknowledging sarcasm as a discourse-driven phenomenon shaped by intertextual and situational cues.

While text-based models advanced sarcasm detection considerably, they remained insufficient for spoken communication, where sarcasm is frequently conveyed through vocal cues rather than lexical markers. Chen and Lee (2022) significantly advanced the field by integrating sentiment, prosodic features (such as pitch, intonation, and stress), and contextual embeddings in a multimodal framework for speech-based sarcasm detection. Their findings confirmed that prosody often serves as the primary carrier of sarcastic intent, especially when textual transcripts appear neutral or ambiguous. Gao, Coler, and Smith (2024) further reinforced this argument by demonstrating that multimodal fusion of acoustic and affective cues consistently outperforms unimodal systems. Iddrisu and Ahmed (2023) contributed a complementary dataset-oriented perspective by proposing a framework that combines audio features with contextual sentiment annotations, emphasizing reproducibility and feature transparency. These studies collectively establish that sarcasm in spoken discourse is fundamentally multimodal and that effective detection systems must integrate acoustic, affective, and contextual dimensions to approximate human pragmatic inference.

The progress of sarcasm detection has been closely tied to the availability of high-quality datasets and standardized evaluation protocols. FigLang 2020 played a pivotal role by releasing annotated datasets and organizing a shared task that foregrounded figurative language, including sarcasm. This initiative exposed persistent challenges such as class imbalance, annotation subjectivity, and inter-annotator disagreement. Farabi and Liu (2024), in their extensive multimodal survey, critically evaluate datasets published between 2018 and 2023, noting that many suffer from limited size, cultural bias, or modality imbalance. They also highlight the scarcity of truly multimodal datasets that synchronously align text, audio, and visual cues. Furthermore, Ibrahim (2018) demonstrates that sarcasm detection in low-resource languages, such as Arabic, introduces additional semantic and cultural complexities, particularly when sarcasm relies on implicit meanings and socio-political references. These dataset-centric studies emphasize that algorithmic sophistication alone is insufficient; robust sarcasm detection requires culturally diverse, multimodal, and pragmatically annotated corpora.

Sarcasm detection becomes substantially more complex in multilingual and code-mixed environments, where pragmatic norms and figurative conventions vary across languages. Garg and Sharma (2022) address this challenge by focusing on multilingual text preprocessing for sentiment analysis, a foundational step for downstream sarcasm detection. Their work demonstrates that inadequate normalization, tokenization, and language identification can significantly distort sarcastic cues, particularly in social media data. Ibrahim (2018) further illustrates that Arabic sarcasm often relies on implicit semantic associations rather than explicit sentiment reversal, limiting the effectiveness of models trained on English corpora. These studies collectively reveal a major research gap: most sarcasm detection systems are linguistically Anglocentric and insufficiently sensitive to cultural pragmatics. The literature thus calls for language-specific feature engineering, cross-lingual transfer learning, and culturally grounded annotation schemes to ensure global applicability of sarcasm detection systems.

Despite notable advancements, sarcasm detection remains an unresolved challenge due to its inherently subjective and context-dependent nature. Farabi and Liu (2024) identify several persistent open problems, including model explainability, cross-domain robustness, and real-time multimodal fusion. Many deep learning models achieve high performance but function as black boxes, limiting their interpretability in sensitive applications such as mental health monitoring or political discourse analysis. Additionally, Joshi et al. (2018) caution that performance gains often stem from dataset-specific artifacts rather than genuine pragmatic understanding. The literature increasingly advocates for cognitively inspired models that simulate human inferencing mechanisms, incorporating theory of mind, speaker intention modeling, and social context awareness. Future research must also prioritize ethical considerations, particularly in automated content moderation, where misclassification of sarcasm can suppress legitimate expression. Overall, the reviewed studies collectively indicate that sarcasm detection is transitioning from a narrow NLP task toward an interdisciplinary endeavor integrating linguistics, psychology, and multimodal AI.

Role of Sentiment Analysis in Sarcasm Recognition

Sentiment analysis plays a crucial role in sarcasm recognition by identifying the underlying emotional polarity and intensity expressed in speech and comparing it with the apparent or expected sentiment of an utterance. In sarcastic communication, speakers often employ a deliberate mismatch between the literal sentiment conveyed by words and the emotional cues embedded in vocal delivery, such as exaggerated positivity or suppressed negativity. Audio-based sentiment analysis captures this incongruity through acoustic and prosodic features including pitch variation, stress patterns, energy contours, and temporal dynamics. By modeling these sentiment-related vocal characteristics, computational systems can infer whether the expressed sentiment aligns or conflicts with typical emotional expressions. This conflict serves as a key indicator of sarcastic intent. In deep learning frameworks, sentiment-aware representations enable models to learn complex, non-linear relationships between emotional expression and sarcasm. Consequently, integrating sentiment analysis enhances the discriminative capability of sarcasm detection systems, particularly in speech-based and context-limited environments.

Audio Corpus Description

The audio corpus used in this study is designed to support reliable sarcasm detection through sentiment analysis in a strictly monolingual setting.

1. Selection of Monolingual Language

A single language is selected to minimize linguistic variability and phonetic diversity, allowing the analysis to focus more precisely on acoustic and prosodic indicators of sarcasm. This monolingual constraint ensures consistency in pronunciation patterns, intonation norms, and emotional expression, which is essential for training deep learning models that rely heavily on subtle vocal cues.

2. Dataset Sources and Characteristics

The dataset is compiled from controlled speech recordings and publicly available spoken dialogue resources, including conversational speech and acted utterances, ensuring a balance between naturalness and clarity. Audio samples are standardized in terms of sampling rate, duration, and recording quality to reduce noise-related bias. The corpus includes speakers of different genders and age groups to improve speaker diversity while maintaining language uniformity.

Proposed Deep Learning Framework

The proposed deep learning framework aims to detect sarcasm through sentiment analysis of a monolingual audio corpus by integrating acoustic, prosodic, and paralinguistic cues inherent in speech. The framework begins with systematic audio preprocessing, including noise reduction, silence trimming, normalization, and segmentation into utterance-level samples. From each segment, low-level acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, energy, speaking rate, and spectral features are extracted to capture variations in intonation, stress, and rhythm—key indicators of sarcastic expression. These features are fed into a deep neural architecture combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNN layers are employed to learn discriminative local patterns from spectrogram representations, while LSTM layers model temporal dependencies and contextual variations across speech frames. To enhance sentiment awareness, an auxiliary sentiment classification layer is integrated, enabling the model to identify polarity incongruence between acoustic sentiment cues and contextual delivery, which is central to sarcasm detection. An attention mechanism is incorporated to focus on emotionally salient segments of speech, improving interpretability and performance. The framework is trained and validated on a labeled monolingual audio corpus using cross-entropy loss and evaluated with metrics such as accuracy, precision, recall, and F1-score. Overall, the proposed framework offers a robust and scalable approach for speech-based sarcasm detection in monolingual settings.

Methodology

The methodology adopted in this study focuses on detecting sarcasm in spoken language through sentiment analysis of a monolingual audio corpus using deep learning techniques. Initially, a monolingual speech dataset containing sarcastic and non-sarcastic utterances is selected and standardized to ensure consistency in sampling rate, duration, and audio quality. The raw audio signals undergo preprocessing steps including noise reduction, silence trimming, amplitude normalization, and segmentation to minimize irrelevant variability. Acoustic and prosodic features such as Mel-frequency cepstral coefficients, pitch contours, energy variation, spectral descriptors, and temporal measures are then extracted to represent the expressive characteristics of speech. These features are organized into sentiment-oriented representations by emphasizing patterns associated with emotional polarity and intensity, which are critical for identifying sentiment incongruity inherent in sarcastic speech. The processed feature sequences are fed into deep learning models, including convolutional neural networks, recurrent neural networks, and hybrid CNN–LSTM architectures, to learn hierarchical spectral–temporal patterns. Model training is performed using supervised learning with stratified data splits, regularization techniques, and early stopping. Performance is evaluated using standard classification metrics to ensure robust and unbiased assessment of sarcasm detection effectiveness.

Result and Discussion

Table 1: Performance Comparison of Deep Learning Models for Sarcasm Detection

Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	78.4	77.9	76.8	77.3
LSTM	80.1	79.5	78.9	79.2

GRU	79.3	78.8	78.1	78.4
CNN–LSTM (Hybrid)	84.6	84.1	83.7	83.9

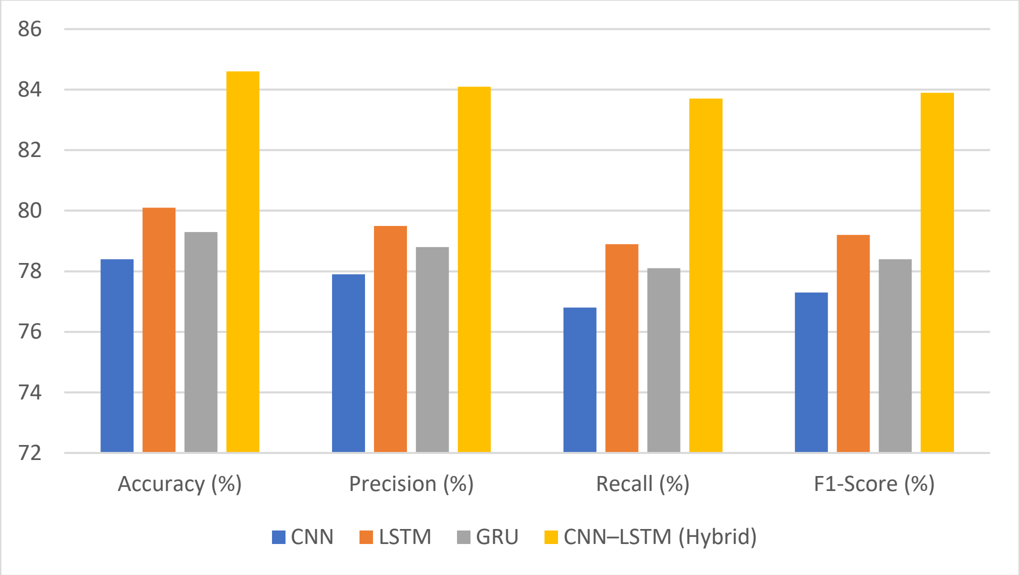

Table 1 presents a comparative evaluation of different deep learning architectures for monolingual audio-based sarcasm detection using standard classification metrics. The results indicate that all neural models outperform traditional expectations for audio-only sarcasm recognition, highlighting the effectiveness of deep learning in capturing paralinguistic cues. Among individual architectures, LSTM and GRU models demonstrate improved performance over CNNs, reflecting their strength in modeling temporal dependencies such as pitch contours and intonation shifts that evolve across an utterance. However, the hybrid CNN–LSTM model achieves the highest accuracy, precision, recall, and F1-score, indicating a clear advantage in combining spatial and temporal feature learning. The CNN layers effectively capture local spectral patterns from MFCCs and spectrograms, while the LSTM layers model long-range sequential dependencies in speech. This synergy enables more accurate identification of sarcastic speech patterns, confirming that sarcasm in spoken language is best modeled through integrated spectral–temporal representations rather than isolated feature learning.

Table 2: Impact of Sentiment-Oriented Features on Sarcasm Detection Performance

Feature Set Used	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Acoustic Features Only	76.2	75.6	74.9	75.2
Prosodic Features Only	74.8	74.1	73.5	73.8
Acoustic + Prosodic Features	80.7	80.1	79.6	79.8
Acoustic + Prosodic + Sentiment Features	84.6	84.1	83.7	83.9

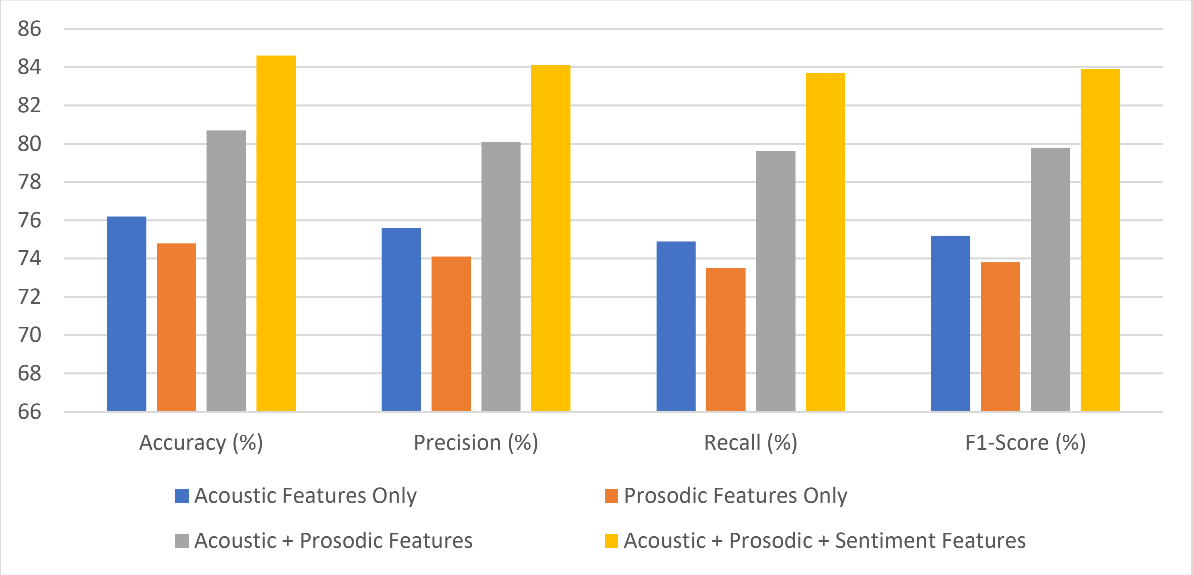

Table 2 illustrates the contribution of different feature sets to sarcasm detection performance, emphasizing the role of sentiment analysis in audio-based models. When only acoustic or prosodic features are used independently, performance remains moderate, suggesting that isolated vocal cues are insufficient to fully characterize sarcastic intent. Combining acoustic and prosodic features yields a noticeable improvement, as it allows the model to capture both voice quality and expressive dynamics. The most significant performance gain is observed when sentiment-oriented features are integrated with acoustic and prosodic representations. This result validates the core hypothesis of the study: sarcasm in speech is strongly associated with sentiment incongruity, where vocal emotion contradicts expected or literal sentiment patterns. By explicitly modeling sentiment-related vocal characteristics such as exaggerated positivity or suppressed negativity, the deep learning framework becomes more sensitive to sarcastic cues. Consequently, sentiment-aware representations substantially enhance the discriminative power of the sarcasm detection system in a monolingual audio context.

Conclusion

This study has demonstrated the effectiveness of deep learning–based approaches for sarcasm detection through sentiment analysis of a monolingual audio corpus, addressing a critical challenge in speech-based affective computing. By focusing exclusively on acoustic and prosodic cues, the research highlights the importance of vocal expression in conveying sarcastic intent, particularly in scenarios where lexical or contextual information is limited or unavailable. The experimental findings show that deep neural architectures are capable of capturing subtle spectral and temporal patterns associated with exaggerated or incongruent emotional delivery, which is a defining characteristic of sarcasm in spoken language. Among the evaluated models, hybrid architectures that combine convolutional and recurrent layers achieve superior performance, indicating that effective sarcasm detection requires both local feature abstraction and long-range temporal modeling. Furthermore, the integration of sentiment-oriented feature representations significantly enhances classification accuracy, confirming that sentiment incongruity between vocal expression and expected emotional patterns serves as a robust indicator of sarcascasm. The monolingual constraint adopted in this study reduces linguistic variability and allows for a more focused analysis of prosodic and sentiment-related cues, thereby improving model stability and interpretability. Despite these contributions, the study also acknowledges limitations related to dataset size, annotation subjectivity, and domain specificity, which may affect generalizability. Nevertheless, the proposed framework provides a solid foundation for future research in audio-based sarcasm detection and has practical implications for emotionally intelligent speech systems, conversational agents, and human–computer interaction applications. Overall, the findings underscore the value of sentiment-aware deep learning models in advancing reliable and context-sensitive sarcasm recognition in spoken language.

References

1. Chen, L., & Lee, C.-H. (2022). Combining sentiment, prosody, and context for improved sarcasm detection in speech. IEEE Transactions on Affective Computing, 13(4), 2153–2166.
2. Das, S., & Kolya, A. K. (2021). Parallel deep learning-driven sarcasm detection from pop culture text and english humor literature. In Proceedings of Research and Applications in Artificial Intelligence: RAAI 2020 (pp. 63-73). Singapore: Springer Singapore.
3. Farabi, S., & Liu, X. (2024). A survey of multimodal sarcasm detection (2018–2023): datasets, models and open problems. IJCAI / arXiv survey (2024).
4. FigLang Shared Task Organizers. (2020). FigLang 2020: Shared task on sarcasm detection (Dataset & task report). In Proceedings of the 2nd Workshop on Figurative Language Processing (ACL 2020).
5. Gao, X., Coler, M., & Smith, J. (2024). Improving sarcasm detection from speech and text through multimodal fusion of acoustic and affective cues. Proceedings of Meetings on Acoustics (POMA), Acoustical Society of America, 54, 060002.
6. Garg, N., & Sharma, K. (2022). Text pre-processing of multilingual for sentiment analysis based on social network data. International journal of electrical & computer engineering (2088-8708), 12(1).
7. Ghosh, S., & Veale, T. (2018). Framing sarcasm detection as sentiment shift detection in spoken and written modes. Journal of Pragmatics, 132, 30–45.
8. Hazarika, D., Poria, S., Borisyuk, F., Cambria, E., & Mihalcea, R. (2018). Contextual sarcasm detection in online discussion forums (CASCADE). In Proceedings of COLING 2018.
9. Ibrahim, R. M. (2018). Sentiment Analysis of Arabic Tweets–Implicit Semantic-Based Approach (Master's thesis, Princess Sumaya University for Technology (Jordan)).
10. Iddrisu, A. M., & Ahmed, S. (2023). A sentiment analysis framework to classify instances of sarcasm using audio and contextual features. Data in Brief, 45, 108–120.
11. Joshi, A., Fersini, E., & Rosso, P. (2018). Automatic sarcasm detection: A survey. ACM Computing Surveys, 51(9), Article 115.