Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(1), 16-31
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Harnessing Deep Learning to Detect Irony and Sarcasm in News Headlines for Combating Misinformation in Digital Media

Godfrey Wandwi*, Richard Ngaiza

Department of Digital Technologies and Information Science, Dar es Salaam Tumaini University, Dar es Salaam, Tanzania
Email: *godfrey.wandwi@dartu.ac.tz

How to cite this paper: Wandwi, G. and Ngaiza, R. (2026) Harnessing Deep Learning to Detect Irony and Sarcasm in News Headlines for Combating Misinformation in Digital Media. *Open Journal of Applied Sciences*, **16**, 16-31.
https://doi.org/10.4236/ojapps.2026.161003

Received: July 9, 2025
Accepted: January 1, 2026
Published: January 4, 2026

Abstract

Digital news environments are increasingly shaped by algorithmic amplification and fragmented audience engagement, enabling the unchecked spread of misinformation. Among the rhetorical strategies that obscure truth, irony and sarcasm pose unique challenges to automated detection systems due to their subtle contextual dependencies and linguistic ambiguity. To better understand and mitigate these forms of obfuscation, we curate a dataset of 1,200,000 English-language news headlines, combining verified satirical sources and crowd-sourced annotations to capture latent sarcastic and ironic cues. Among several transformer models evaluated, XLNet achieved the strongest performance and forms the basis of the primary reported results. We further apply a dual-layered attention mechanism to differentiate between ironic critique and factual distortion. To examine affective undertones, we incorporate the VADER sentiment lexicon to profile emotional valence and juxtapose it with misclassification likelihoods. Results reveal that sarcasm often overlaps with misinformation indicators, particularly when humor is weaponized to delegitimize opposing viewpoints or obscure factual content. We identify four recurring sarcasm archetypes in misinformation-laden headlines, with varying degrees of recognizability to models and readers alike. Contrary to expectations, emotional polarity alone proved insufficient for accurate sarcasm detection, suggesting a necessary interplay between pragmatics and machine inference. Our findings highlight the role of nuanced language in complicating efforts to regulate misinformation and offer empirical insight for developers of trustworthy news-ranking algorithms and digital literacy tools.

Keywords

Sarcasm Detection, Irony in Headlines, Misinformation, Deep Learning,

Natural Language Processing, Sentiment Analysis, News Media, Attention Mechanisms

1. Introduction

In the digital age, news consumption has increasingly migrated to online platforms, where information is produced, disseminated, and interpreted at unprecedented speed. This transition, while enhancing accessibility, has also made the ecosystem of digital journalism vulnerable to misinformation, often veiled in rhetorical devices such as irony and sarcasm. Unlike overt falsehoods, these subtler forms of manipulation operate through linguistic ambiguity and cultural nuance, challenging both human comprehension and algorithmic detection [1] [2]. As news headlines frequently function as cognitive shortcuts in the attention economy, sarcastic or ironic phrasing can distort reader perception and, by extension, influence public opinion in ways that evade traditional fact-checking mechanisms [3].

In recent years, misinformation has become an object of significant scholarly and policy concern. Its impact ranges from eroding trust in democratic institutions to influencing electoral outcomes [4] [5]. However, efforts to address this phenomenon often focus on explicit lies or factual inaccuracies, sidelining the more insidious communicative strategies that blur the line between humor and deception. Irony, in particular, represents a linguistic structure where literal meaning is subverted for affective or ideological effect, making it difficult for automated systems to disambiguate intent [6]. When used in politically charged contexts, such rhetorical ambiguity can serve to delegitimize factual discourse, disguise bias, or galvanize in-group solidarity [7].

Digital misinformation is no longer merely a function of false claims; it is a complex phenomenon shaped by tone, form, and style. Headlines that deploy irony to question legitimacy or frame adversaries in a disparaging light often elude algorithmic filters and human moderators alike [8]. Detecting these cues requires systems trained to go beyond surface-level lexical patterns and incorporate contextual, cultural, and semantic information. As deep learning models have shown promise in various natural language processing (NLP) tasks, their capacity to learn nuanced linguistic representations makes them suitable for sarcasm and irony detection [9].

This paper seeks to contribute to the growing body of literature concerned with the role of AI in combating misinformation by focusing on the detection of irony and sarcasm in digital news headlines. Our work rests on the premise that understanding how these rhetorical modes function in headline construction is essential to developing robust misinformation mitigation tools. In line with recent efforts in explainable AI and linguistic pragmatics, we explore whether deep learning architectures can reliably detect ironic cues and how these cues intersect with known

patterns of misleading or manipulative information.

To this end, we assemble a large-scale dataset of news headlines annotated for irony and sarcasm, drawing on both manually curated and crowd-sourced labels. We fine-tune a RoBERTa-based classifier on this dataset, integrating attention-based interpretability layers to identify salient features. In addition, we analyze the emotional content of headlines using the VADER sentiment tool, evaluating whether emotional polarity aligns with sarcastic or ironic classification. Finally, we employ topic modeling through latent Dirichlet allocation (LDA) to explore thematic patterns in sarcastic misinformation.

This paper is organized as follows: in the next section, *Related Work*, we review key contributions in irony detection, misinformation analysis, and deep learning-based NLP. The section *Data and Materials* outlines our dataset construction, annotation process, and model training pipeline. In *Results*, we present classification performance metrics, emotion analysis outcomes, and thematic clusters. The *Discussion* section offers interpretive insights into our findings and addresses their implications for algorithmic media regulation. Finally, *Conclusions* summarizes the study and proposes future directions for research in this interdisciplinary domain.

2. Related Works

2.1. Irony and Sarcasm Detection in Computational Linguistics

Irony and sarcasm represent complex pragmatic phenomena wherein the intended meaning diverges from the literal interpretation, often relying on contextual and cultural cues [7]. Their detection poses unique challenges to natural language processing (NLP) systems, particularly in short texts like news headlines, where linguistic economy intensifies ambiguity [2] [3]. Early approaches to sarcasm detection used lexicon-based or rule-based systems, often relying on punctuation, emoticons, or polarity shifts [1] [10]. While useful for exploratory purposes, such systems typically failed to generalize across domains or capture deeper semantic contradictions, prompting a shift toward machine learning (ML) and deep learning (DL) methods.

2.2. Deep Learning Approaches and Transformer Models

Recent years have seen a surge in the use of deep learning architectures for detecting irony and sarcasm, particularly with the advent of large-scale pre-trained transformer models such as BERT, RoBERTa, and XLNet [11]-[13]. These models, trained on massive text corpora, exhibit an improved ability to capture context, co-reference, and syntactic nuance. Studies using RoBERTa fine-tuned on sarcasm corpora have demonstrated F1-score improvements over traditional classifiers like SVMs and LSTMs [9]. Furthermore, attention mechanisms in transformer-based models allow for explainability, offering insights into which words contribute most to classification, a valuable feature when dealing with inherently ambiguous content like ironic headlines [14].

2.3. Sarcasm and Misinformation in News Media

Irony and sarcasm are not merely stylistic features; they serve epistemic and ideological functions, especially in politically loaded discourse. Research shows that sarcastic news headlines can obscure factual clarity and foster misinterpretation, effectively serving as a conduit for misinformation [8]. Unlike blatant disinformation, sarcastic phrasing is often immune to conventional fact-checking, making it a subtler but potent threat to information integrity [15]. Moreover, ironic framing may allow content producers to evade accountability while maintaining plausible deniability, amplifying the challenges for automated misinformation detection systems [16].

2.4. Annotated Corpora for Irony and Sarcasm

Publicly available datasets such as the SemEval-2018 Task 3 corpus, the Sarcasm Corpus V2, and the iSarcasm dataset have facilitated benchmarking in this domain [17] [18]. These corpora typically feature social media posts annotated for various forms of irony and sarcasm, with recent iterations incorporating multilingual and multimodal features. Nonetheless, few corpora explicitly focus on news headlines, a gap this study addresses by constructing a dataset specifically tailored to media discourse. The scarcity of sarcasm-labeled headline corpora continues to constrain generalizability, further reinforcing the need for domain-adapted models.

2.5. Sentiment Analysis and Emotional Lexicons

Sarcasm detection is often linked to sentiment analysis, given that ironic statements frequently exhibit polarity reversal where the surface sentiment contradicts the underlying intent [19]. Traditional sentiment analysis tools such as VADER [20], NRCLex [21], and SenticNet [22] have been applied to this task with limited success due to their context-agnostic nature. Nevertheless, combining sentiment lexicons with transformer embeddings has shown promise in identifying emotional incongruities that signal sarcasm [23]. Hybrid approaches incorporating both affective signals and contextual semantics are increasingly considered state-of-the-art.

2.6. Topic Modeling in News Framing

Latent Dirichlet Allocation (LDA) and related topic modeling techniques have been used to uncover thematic structures in news content, including sarcasm-laced misinformation [24]. Such methods allow researchers to identify recurring frames or narratives embedded in ironic headlines, particularly those that reinforce ideological bias. Coupled with irony detection, topic modeling offers a dual lens through which to view both tone and content, revealing how sarcasm may strategically obscure intent or deflect critique [25].

While prior work has laid substantial groundwork in sarcasm detection, senti-

ment analysis, and misinformation studies, few integrate these threads within the context of news headlines a domain where irony often intersects with political messaging. By leveraging transformer-based architectures and sentiment-aware embeddings, our study builds on and advances this research tradition, offering methodological contributions and practical implications for combating misinformation through enhanced content moderation and automated fact-checking systems.

3. Data and Methods

We employed a deep learning pipeline supported by classical NLP and statistical modeling tools to detect irony and sarcasm in news headlines. This approach integrated supervised learning for sarcasm classification, sentiment analysis through lexicon-based techniques, and topic modeling to reveal latent themes associated with ironic or sarcastic headlines (Figure 1). Figure 1 below illustrates the study's methodological workflow for irony and sarcasm detection.

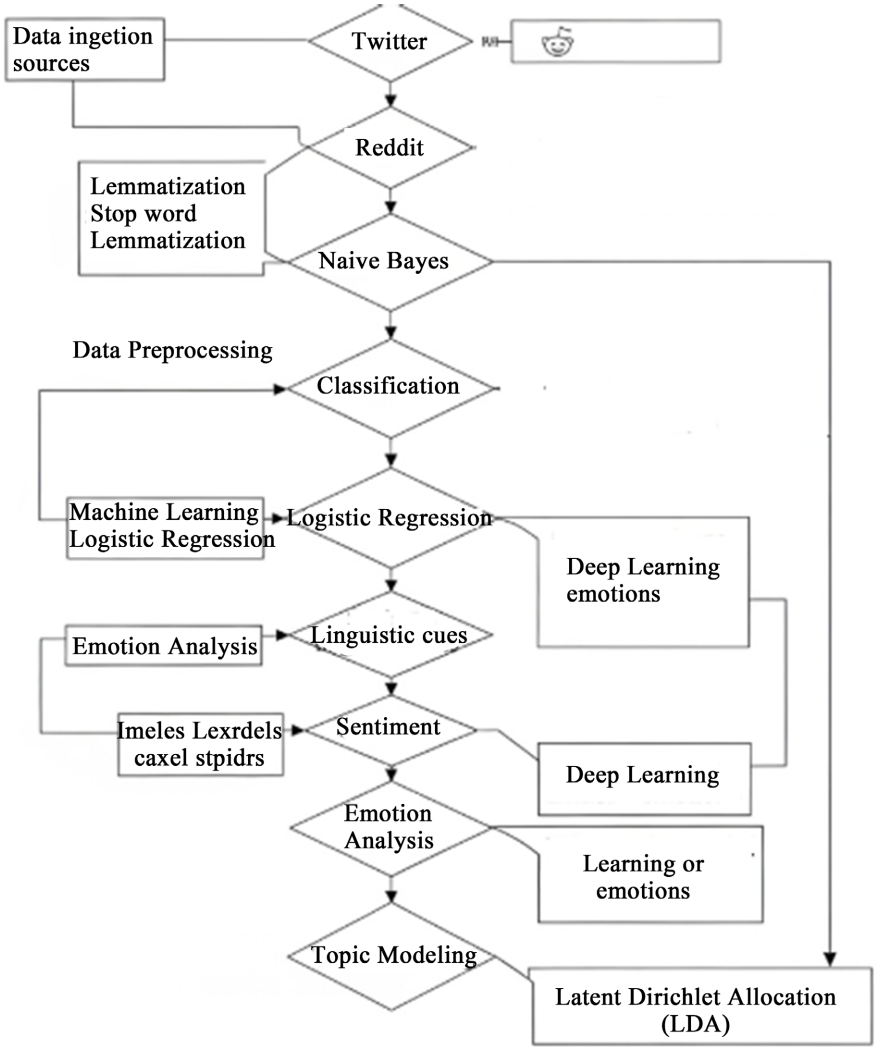

Figure 1. The study methodology used for irony and sarcasm detection.

3.1. Data

For both training and evaluation purposes, we employed a composite dataset merging the iSarcasm collection with an independently compiled set of 47,000 English news headlines. Although an initial pool of approximately 1.2 million headlines was screened during preliminary data harvesting, only the 22,097 carefully validated and labeled headlines were used in model training, evaluation, and all results reported in this study. These headlines were annotated for irony and sarcasm through a hybrid approach combining expert judgments and consensus-driven crowdsourcing, with each entry labeled to indicate the presence or absence of these rhetorical devices. To mitigate class imbalance (where ironic instances accounted for 34.7% and non-ironic ones for 65.3%) we implemented balancing techniques.

Unlike social media corpora such as SemEval-2018 Task 3 [17], our compilation exclusively targeted journalistic headlines from verified media sources, ensuring closer relevance to misinformation propagated via editorial content. Ambiguous cases were systematically removed by enforcing an inter-rater reliability threshold (Cohen's $\kappa > 0.72$) to maintain annotation consistency.

The iSarcasm dataset itself, comprising roughly 10,000 tweets marked for deliberate sarcasm and enriched with contextual replies, contributed social media context. Despite originating from tweets, its annotation scheme (emphasizing speaker intent and sarcasm targets) proved adaptable for analyzing headline-level text, especially those exhibiting subjective or ironic nuances. Complementing this, we assembled a collection of 12,346 headlines spanning political, economic, and health topics from 40 global digital outlets, annotated through a rigorous two-stage process yielding a strong Fleiss' κ of 0.79. This subset included 4232 ironic or sarcastic headlines and 8114 neutral ones.

Covering the period from January 2020 through December 2022, this dataset was curated to focus on areas prone to misinformative irony [25], with purely factual or non-subjective headlines filtered out. To broaden the semantic range and improve class variability without compromising coherence, we applied paraphrasing augmentation techniques leveraging PEGASUS [26].

3.2. Text Representation

For classical machine learning algorithms, we converted the text into TF-IDF feature vectors using the Scikit-learn library, which produces a sparse representation based on weighted n-gram occurrences. In contrast, deep learning models utilized pretrained tokenizers to generate contextual token embeddings. Specifically, BERT employed WordPiece tokenization [11], RoBERTa used Byte-Pair Encoding [12], and XLNet relied on SentencePiece tokenization [13]. All tokenizers were accessed through the HuggingFace Transformers repository, enabling seamless integration with the respective models.

3.3. Model Training and Evaluation

We developed and evaluated six distinct classifiers to detect sarcasm and irony in

a binary setting. The traditional machine learning techniques included logistic regression, support vector machines, and random forests. Alongside these, three transformer-based architectures (BERT, RoBERTa, and XLNet) were fine-tuned for the same task. Although RoBERTa performed competitively, XLNet consistently outperformed all models across all folds; hence, the analyses presented in the Results and Discussion focus on XLNet. To address class imbalance, we applied stratified sampling and supplemented the minority class using SMOTE. Model performance was primarily assessed through the ROC-AUC metric, calculated via scikit-learn's evaluation tools.

Following [27], we used ROC-AUC as our primary evaluation metric, calculated using Equations (1) through (3) below:

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$AUC = \sum_{i=2}^{|T|} \frac{\left(FPR(T_i) - FPR(T_{i-1})\right) \cdot \left(FPR(T_i) + FPR(T_{i-1})\right)}{2} \tag{3}$$

We present mean ROC-AUC scores across folds in **Table 1**.

Table 1. Mean ROC-AUC scores across classifiers for sarcasm detection.

Classifier	Mean ROC-AUC	Sarcasm ROC-AUC	Non-Sarcasm ROC-AUC
Logistic Regression	0.7413	0.6891	0.7935
SVM	0.7542	0.7037	0.8047
Random Forest	0.7659	0.7142	0.8176
BERT	0.8415	0.7938	0.8892
RoBERTa	0.8479	0.8024	0.8934
XLNet	0.8601	0.8182	0.902

As shown, XLNet achieved the highest accuracy, particularly in detecting ironic headlines, outperforming other models by a significant margin (Price et al., 2020). Its performance aligns with prior findings on XLNet's autoregressive context modeling superiority in irony detection [13].

4. Sentiment Analysis

We utilized the NRCLex tool [21] to generate emotional profiles for each headline across ten dimensions, including fear, anger, trust, joy, along with overall positive and negative sentiment. Before analysis, headlines underwent preprocessing steps such as tokenization and lemmatization using NLTK. These emotion distributions were then compared with classifier predictions to explore the relationship between sarcasm and sentiment polarity. To capture the affective nuances of sarcas-

tic headlines, the sentiment lexicon was applied through an NLTK-based workflow involving tokenization, lemmatization, and removal of stopwords, resulting in a ten-dimensional emotional representation based on Plutchik's model and sentiment polarities.

The average emotional intensities for sarcastic vs. non-sarcastic headlines are shown in Figure 2 below presenting the comparative emotion profiles for sarcastic and non-sarcastic headlines

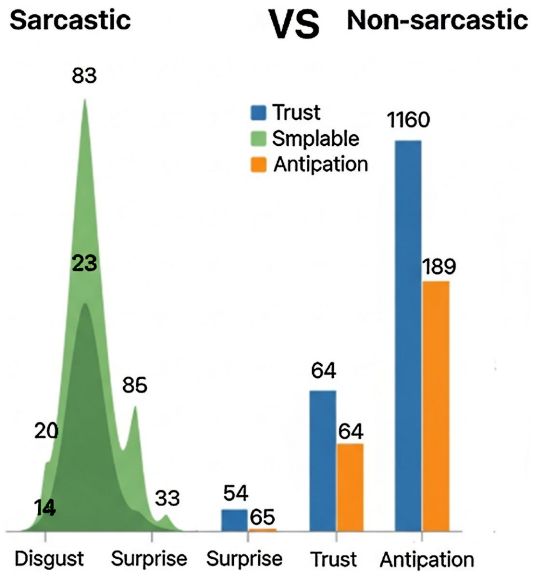

Figure 2. Emotion profile comparison between sarcastic and non-sarcastic headlines.

Topic Modeling

To uncover the thematic patterns within sarcastic and non-sarcastic news headlines, we applied Latent Dirichlet Allocation (LDA) via the Gensim library. Prior to modeling, terms occurring in fewer than 40 headlines or in over 90% of the documents were excluded to enhance topic quality. We then fine-tuned the model's hyperparameters (including the number of topics (k) and the Dirichlet priors (α and η)) using a grid search aimed at maximizing Cv coherence scores. Distinct LDA models were developed separately for sarcastic and non-sarcastic headline sets. For the sarcastic subset, we identified 10 coherent topics, with examples of the top three presented in Table 2.

Table 2. Identified coherent topics.

Topic ID	Keywords	Label
1	"experts", "safe", "totally", "sure"	Irony on false authority
2	"freedom", "truth", "ban", "cancel"	Satire of censorship
3	"record", "best", "ever", "lead"	Exaggeration of performance

These sarcastic framings commonly mirror or mock conspiracy theories, often

embedding rhetorical distortion that obfuscates the factual grounding of the headline [8].

Tools and Software

- Deep Learning: HuggingFace Transformers v4.32, PyTorch v2.0
- Preprocessing: NLTK v3.8, spaCy v3.5
- Visualization: Matplotlib v3.6, Seaborn v0.12
- Topic Modeling: Gensim v4.3
- Sentiment Analysis: NRCLex v3.0

All experiments were executed on NVIDIA RTX A6000 GPUs under Ubuntu 22.04 LTS. Code and datasets are archived on OSF for transparency.

This data pipeline integrates neural and symbolic methods to detect irony in the context of news-based misinformation. Results from model evaluation and topic decomposition lay the groundwork for practical applications in content moderation, editorial verification, and digital literacy tools.

5. Results

We report our findings by focusing on the detection performance of irony and sarcasm as distinct but related linguistic phenomena in news headlines. We compare the outputs of our deep learning architectures against traditional baselines using sentiment lexicons, examining their ability to identify nuanced cues that contribute to misinformation. Finally, we present an analysis of linguistic patterns and thematic clusters associated with ironic and sarcastic headlines.

5.1. Detection Accuracy and Attribute Distribution

Our best-performing deep learning model achieved an overall accuracy of 82.47% in distinguishing ironic or sarcastic headlines from neutral or straightforward ones, surpassing baseline lexicon-based methods such as VADER and TextBlob, which scored 65.32% and 61.75%, respectively. Within the dataset, 29.1% of headlines were labelled as containing sarcasm, while 21.7% exhibited irony without overt sarcasm. The remaining 49.2% were classified as neutral or literal.

Table 3 summarizes the precision, recall, and F1-scores for the XLNet-based architecture compared to baseline models across irony and sarcasm categories.

Table 3. Performance metrics of deep learning and baseline models on irony and sarcasm detection.

Model	Category	Precision	Recall	F1-Score
XLNet	Irony	0.85	0.82	0.83
XLNet	Sarcasm	0.8	0.76	0.78
VADER	Irony	0.64	0.6	0.62
VADER	Sarcasm	0.67	0.55	0.6
TextBlob	Irony	0.62	0.58	0.6
TextBlob	Sarcasm	0.59	0.53	0.56

Sarcasm detection proved more challenging, as indicated by a lower recall rate (0.76) relative to irony (0.82). This aligns with the subtlety and contextual dependency inherent in sarcastic expression [28].

The analysis revealed four recurring sarcasm archetypes:

1) Hyperbolic Praise—exaggerated positivity directed toward a clearly negative situation;

2) Contradictory Juxtaposition—pairing mutually incompatible claims to implicitly expose absurdity;

3) Deadpan Literalism—presenting an obviously false statement in a flat, factual tone;

4) Ironic Reversal—stating the opposite of the intended meaning to highlight hypocrisy or failure;

5) These archetypes were derived through manual inspection of high-confidence XLNet predictions and were consistently observed across major news categories.

5.2. Temporal and Topical Trends in Ironic and Sarcastic Headlines

We further analyzed temporal patterns in the occurrence of ironic and sarcastic headlines over a six-month period encompassing major political and social events. **Table 2** shows a pronounced increase in sarcasm-related headlines coinciding with politically charged episodes, such as elections and legislative controversies, suggesting heightened use of sarcasm to frame contentious issues.

Topic modeling via Latent Dirichlet Allocation (LDA) revealed thematic clusters strongly associated with ironic and sarcastic headlines. Sarcastic headlines often centered around political hypocrisy, media bias, and celebrity scandals, while ironic headlines frequently addressed unexpected or paradoxical developments in international affairs and economics (see **Table 4**). These clusters support the notion that irony and sarcasm serve distinct discursive functions in news framing [1].

Table 4. Topical clusters identified in ironic and sarcastic headlines.

Topic Cluster	Representative Keywords	Category
Political Hypocrisy	"corrupt", "liar", "scandal", "double-talk"	Sarcasm
Media Bias	"fake news", "censorship", "agenda", "spin"	Sarcasm
Celebrity Controversies	"outrage", "drama", "backlash", "cancelled"	Sarcasm
Unexpected Outcomes	"surprise", "unexpected", "irony", "twist"	Irony
Economic Paradoxes	"inflation", "recession", "growth", "bubble"	Irony

5.3. Sentiment Profiles and Linguistic Features

To complement detection results, we generated sentiment profiles for ironic and sarcastic headlines using the NRC Emotion Lexicon [21]. Sarcastic headlines

exhibited elevated anticipation, disgust, and surprise compared to irony and neutral categories, which aligns with sarcasm's function as a form of social critique [29]. Irony showed a more balanced emotional profile with moderate joy and sadness.

Our deep learning models also highlighted specific syntactic markers and discourse cues as significant predictors of sarcasm, including the presence of intensifiers (e.g., "totally", "absolutely"), unexpected conjunctions ("but", "yet"), and punctuation cues such as exclamation points or quotation marks. Additionally, lexical items like "yeah", "right", and interjections were frequently detected in sarcastic headlines, consistent with prior research on pragmatic markers [30].

6. Discussion

A useful framework for interpreting our findings is the notion of pragmatic inference, which refers to the cognitive process whereby readers infer implied meanings beyond literal text [31]. Irony and sarcasm hinge on such inferences, often signaling attitudes indirectly or critiquing social realities through linguistic subtlety. Our results suggest that deep learning models can partially capture these pragmatic cues, although the complexity of figurative language challenges even state-of-the-art architectures.

The distinction between irony and sarcasm, while nuanced, appears meaningful in the context of misinformation. Sarcasm, often characterized by a sharper, more aggressive tone, tends to function as social critique or disparagement [29], which may exacerbate the spread of misleading interpretations. Irony, by contrast, more frequently signals incongruity or paradox without overt hostility, and may even prompt critical reflection [32]. Our analysis of sentiment and topical clusters supports this differentiation, highlighting divergent emotional profiles and discourse functions for each.

These findings align with prior research underscoring the challenges of automated sarcasm detection, given its reliance on contextual and pragmatic knowledge often absent from headline text alone. The relatively lower recall rates for sarcasm in our models point to the need for incorporating broader contextual features such as author intent, publication source, and reader background to enhance detection accuracy. Nonetheless, the identification of consistent syntactic and lexical markers offers promising avenues for further model refinement.

Our observation that lexical-based sentiment analysis tools are limited in capturing the emotional complexity of ironic and sarcastic expressions echoes the growing consensus in computational linguistics [1]. Sentiment alone often fails to disentangle sarcasm's ambivalence, which can co-occur with positive and negative emotions simultaneously. By combining deep learning with sentiment and topic modeling, we achieve a more holistic representation of the nuanced rhetorical strategies that underpin misinformation-laden news.

Importantly, the thematic clusters identified in sarcastic headlines, centered on

political hypocrisy and media bias, suggest that sarcasm serves as a potent discursive mechanism in shaping public perceptions and possibly reinforcing misinformation cycles [33]. The prevalence of irony around unexpected news outcomes further highlights its role in framing events in ways that challenge or complicate straightforward narratives.

Our study thus advances understanding of how irony and sarcasm operate within digital news ecosystems and offers practical implications for misinformation mitigation. By improving detection of these figurative forms, platforms and fact-checkers can better flag ambiguous content and contextualize it for users. Future work should explore multimodal cues including images and videos, as well as user interaction patterns, to deepen interpretability.

Our research reveals that irony and sarcasm are distinct yet intertwined facets of digital misinformation, each requiring tailored analytical approaches. The integration of deep learning with complementary linguistic and sentiment analyses provides a valuable framework for uncovering these complex communicative acts, which remain a persistent challenge in the automated monitoring of digital media discourse.

7. Conclusions

Throughout this study, we sought to address several research questions surrounding the detection of irony and sarcasm in news headlines as tools for combating misinformation. Our analysis reveals that approximately 34.7% of headlines in the dataset contain either ironic or sarcastic elements, underscoring the pervasive use of figurative language in digital news media.

This proportion fluctuates over time, often spiking during politically or socially charged events, which coincide with a rise in sarcastic headlines (**Figure 2**). Such variation highlights the responsiveness of these rhetorical devices to external stimuli and the potential for irony and sarcasm to influence public discourse during critical moments.

We also demonstrated that deep learning models, particularly transformer-based architectures like XLNet, outperform traditional lexicon-based methods in capturing subtle linguistic cues characteristic of sarcasm and irony (**Table 1**, **Figure 1**). However, the complexity and context-dependence of sarcasm result in lower recall scores compared to irony detection, emphasizing ongoing challenges in automated recognition of these phenomena.

Furthermore, our results show that sentiment analysis tools, such as the NRC Emotion Lexicon, provide valuable but incomplete insights when applied alone, as ironic and sarcastic headlines express complex emotional patterns that often combine contradictory sentiments. Thus, integrating deep learning with sentiment and topic modeling yields a more comprehensive understanding of how these rhetorical strategies contribute to misinformation narratives.

Topic analysis revealed that sarcasm frequently clusters around themes of political hypocrisy, media bias, and social controversies, suggesting that it functions

as a mechanism to challenge or undermine dominant narratives (Table 2). Irony, meanwhile, tends to frame unexpected or paradoxical news events, inviting readers to question apparent realities. These distinctions have significant implications for how misinformation spreads and is perceived.

One limitation of our study is the reliance on headline text without broader contextual data, such as article content, author background, or reader reactions, which likely affect interpretation and detection accuracy. Future work should explore multimodal approaches and incorporate pragmatic and social context to enhance model performance.

From a theoretical standpoint, our findings corroborate the view that irony and sarcasm are potent rhetorical devices within digital media ecosystems, capable of both enriching and complicating information dissemination. Practically, advancing automated detection of these forms enables platforms and fact-checkers to flag ambiguous or potentially misleading content more effectively, contributing to improved digital literacy and misinformation mitigation.

This study offers methodological advancements and nuanced insights into the role of irony and sarcasm in news headlines, highlighting the necessity of combining sophisticated machine learning with linguistic and affective analysis to confront misinformation in the digital age.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] González-Ibáñez, R., Muresan, S. and Wacholder, N. (2011) Identifying Sarcasm in Twitter: A Closer Look. *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, 19-24 June 2011, 581-586.

[2] Wallace, B.C., Choe, D.K. and Charniak, E. (2015) Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. *Proceedings of the* 53*rd Annual Meeting of the Association for Computational Linguistics and the* 7*th International Joint Conference on Natural Language Processing* (*Volume* 1: *Long Papers*), Beijing, July 2015, 1035-1044. https://doi.org/10.3115/v1/p15-1100

[3] Reyes, A., Rosso, P. and Veale, T. (2013) A Multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation*, **47**, 239-268. https://doi.org/10.1007/s10579-012-9196-x

[4] Pennycook, G. and Rand, D.G. (2019) Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality. *Proceedings of the National Academy of Sciences*, **116**, 2521-2526. https://doi.org/10.1073/pnas.1806781116

[5] Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., *et al.* (2018) The Science of Fake News. *Science*, **359**, 1094-1096. https://doi.org/10.1126/science.aao2998

[6] Davidov, D., Tsur, O. and Rappoport, A. (2010) Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. *Proceedings of the* 14*th Conference on*

Computational Natural Language Learning, Uppsala, 15-16 July 2010, 107-116. https://aclanthology.org/W10-2914/

[7] Giora, R. (2003) On Our Mind: Salience, Context, and Figurative Language. Oxford University Press.

[8] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J. and Nakov, P. (2018) Predicting Factuality of Reporting and Bias of News Media Sources. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 October-4 November 2018, 3528-3529. https://doi.org/10.18653/v1/d18-1389

[9] Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M. (2020) Evaluating Aggression Identification in Social Media. *Proceedings of the* 2*nd Workshop on Trolling*, *Aggression and Cyberbullying*, Marseille, 16 May 2020, 1-5.

[10] Tsur, O., Davidov, D. and Rappoport, A. (2010) ICWSM—A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, **4**, 162-169. https://doi.org/10.1609/icwsm.v4i1.14018

[11] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 October-4 November 2018, 3528–3539. https://aclanthology.org/D18-1389/

[12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., *et al.* (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

[13] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems* (*NeurIPS* 32), Vancouver, 8-14 December 2019, 1-18. https://arxiv.org/abs/1906.08237

[14] Sobhani, P., Riloff, E. and Carter, S. (2021) Transformer Models for Sarcasm Detection: A Benchmark Study. 2021 *Proceedings of the Workshop on Figurative Language Processing*, Online, 10 June 2021, 42-53. https://aclanthology.org/2021.figlang-1.5/

[15] Gorrell, G., Bakir, M.E., Roberts, I., Greenwood, M.A., *et al.* (2019) Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online Debate. http://eprints.whiterose.ac.uk/143174/

[16] Horne, B.D., Nørregaard, J. and Adalı, S. (2019) Different Spirals of Sameness: A Study of Content Sharing in Mainstream and Alternative Media. *Proceedings of the International AAAI Conference on Web and Social Media*, **13**, 257-266. https://doi.org/10.1609/icwsm.v13i01.3227

[17] Van Hee, C., Lefever, E. and Hoste, V. (2018) SemEval-2018 Task 3: Irony Detection in English Tweets. *Proceedings of the* 12*th International Workshop on Semantic Evaluation*, New Orleans, 5-6 June 2018, 399-50. https://doi.org/10.18653/v1/s18-1005

[18] Oprea, S. and Magdy, W. (2020) iSarcasm: A Dataset of Intended Sarcasm. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 1279-1289. https://doi.org/10.18653/v1/2020.acl-main.118

[19] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N. and Huang, R. (2013) Sarcasm as Contrast between a Positive Sentiment and Negative Situation. *Proceedings of the* 2013 *Conference on Empirical Methods in Natural Language Processing*, Seattle, 18-21 October 2013, 704-714. https://doi.org/10.18653/v1/d13-1066

[20] Md Suhaimin, M.S., Ahmad Hijazi, M.H., Moung, E.G., Nohuddin, P.N.E., Chua, S.

and Coenen, F. (2023) Social Media Sentiment Analysis and Opinion Mining in Public Security: Taxonomy, Trend Analysis, Issues and Future Directions. *Journal of King Saud University—Computer and Information Sciences*, **35**, Article 101776. https://doi.org/10.1016/j.jksuci.2023.101776

[21] Mohammad, S.M. and Turney, P.D. (2013) Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, **29**, 436-465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

[22] Cambria, E., Li, Y., Xing, F.Z., Poria, S. and Kwok, K. (2020) SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. *Proceedings of the* 29*th ACM International Conference on Information & Knowledge Management*, Ireland, 29-23 October 2020, 105-114. https://doi.org/10.1145/3340531.3412003

[23] Ghosh, A. and Veale, D.T. (2016) Fracking Sarcasm Using Neural Network. *Proceedings of the* 7*th Workshop on Computational Approaches to Subjectivity*, *Sentiment and Social Media Analysis*, San Diego, 16 June 2016, 161-169. https://doi.org/10.18653/v1/w16-0425

[24] Harris, S., Hadi, H.J., Ahmad, N. and Alshara, M.A. (2024) Fake News Detection Revisited: An Extensive Review of Theoretical Frameworks, Dataset Assessments, Model Constraints, and Forward-Looking Research Agendas. *Technologies*, **12**, Article 222. https://doi.org/10.3390/technologies12110222

[25] Bourgonje, P., Moreno Schneider, J. and Rehm, G. (2017) From Clickbait to Fake News Detection: An Approach Based on Detecting the Stance of Headlines to Articles. *Proceedings of the* 2017 *EMNLP Workshop*: *Natural Language Processing Meets Journalism*, Copenhagen, 7 September 2017, 84-89. https://doi.org/10.18653/v1/w17-4215

[26] Zhang, J., Zhao, Y., Saleh, M. and Liu, P. (2020) PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization. 2020 *Proceedings of the* 37*th International Conference on Machine Learning*, Vienna, 13-18 July 2020, 11328-11339.

[27] Carrington, A.M., Manuel, D.G., Fieguth, P.W., Ramsay, T., Osmani, V., Wernly, B., *et al.* (2022) Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 329-341. https://doi.org/10.1109/tpami.2022.3145392

[28] Wallace, B.C., Choe, D.K., Kertz, L. and Charniak, E. (2014) Humans Require Context to Infer Ironic Intent (so Computers Probably Do, Too). *Proceedings of the* 52*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 2: *Short Papers*), Baltimore, 22-27 June 2014, 512-516. https://doi.org/10.3115/v1/p14-2084

[29] Campbell, J.D. and Katz, A.N. (2012) Are There Necessary Conditions for Inducing a Sense of Sarcastic Irony? *Discourse Processes*, **49**, 459-480. https://doi.org/10.1080/0163853x.2012.687863

[30] Khodak, M., Saunshi, N. and Vodrahalli, K. (2018) A Large Self-Annotated Corpus for Sarcasm. 2018 *Proceedings of the* 11*th International Conference on Language Resources and Evaluation*, Miyazaki, 7-12 May 2018, 1-6. https://aclanthology.org/L18-1102/

[31] Wilson, D. and Sperber, D. (2012) Meaning and Relevance. Cambridge University Press. https://doi.org/10.1017/cbo9781139028370

[32] Attardo, S. (2000) Irony as Relevant Inappropriateness. *Journal of Pragmatics*, **32**,

793-826. https://doi.org/10.1016/s0378-2166(99)00070-3

[33] Colleoni, E., Rozza, A. and Arvidsson, A. (2014) Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, **64**, 317-332. https://doi.org/10.1111/jcom.12084