Journal Title
XX(X):1–29
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A Systematic Review of Figurative Speech Detection: Methods, Challenges, and Multilingual Perspectives

Zouheir Banou[1], Sanaa Elfilali[1], El Habib Benlahmar[1], Fatima-Zahra Alaoui[1], Hasnae Sakhi Banou[1] and Laila Eljiani[1]

Abstract
Figurative speech detection has emerged as a critical task in natural language processing (NLP), enabling machines to comprehend non-literal expressions such as metaphor, irony, and sarcasm. This study presents a systematic literature review with a multilevel analytical framework, examining figurative language across lexical, syntactic, semantic, discourse, and pragmatic levels. We investigate the interplay between feature engineering, model architectures, and annotation strategies across different languages, analyzing datasets, linguistic resources, and evaluation metrics. Special attention is given to the challenges posed by morphologically rich and low-resource languages, where deep learning dominates but rule-based and hybrid approaches remain relevant. Additionally, we discuss methodological trends, limitations, and future research directions, emphasizing the need for multimodal integration and explainable AI techniques. By structuring our analysis through linguistic and computational levels, this review aims to facilitate the development of more robust and inclusive figurative speech detection systems.

Keywords
Figurative Speech Detection, Natural Language Processing, Literature review, Deep Learning

Introduction

Natural Language Processing (NLP) has advanced significantly in recent years, contributing to tasks such as sentiment analysis, creative writing, and social media analysis. However, figurative speech detection remains an underexplored area, despite its importance in understanding non-literal language, including metaphors, similes, and synecdoche.

Figurative expressions often carry meanings beyond their literal interpretations, making them challenging to detect computationally. The complexity arises from linguistic ambiguity, cultural variations, and the need to model abstract relationships between concepts effectively. Despite these challenges, progress in figurative speech detection is essential for improving semantic understanding and enabling applications in domains such as literary analysis and human-computer interaction.

This review evaluates existing approaches to figurative speech detection. It aims to consolidate findings on datasets, preprocessing techniques, and computational models, identifying gaps and opportunities for future research. By addressing these challenges, this review seeks to advance the development of robust tools for processing figurative language.

The remainder of this paper is organized as follows: section provides the necessary background, detailing the distinctions between these categories and their relevance to figurative speech detection. section describes the methodology employed, including the categorization of figurative styles into lexical, syntactic, and semantic levels. section offers an overview of related work, organized according to these processing levels, while section synthesizes the key findings. Finally, section concludes with recommendations for future research.

Figurative speech and NLP Task levels

NLP faces a unique set of challenges when it comes to analyzing and processing figures of speech. Each linguistic level presents distinct difficulties, ranging from understanding the semantics of individual words to interpreting pragmatic and discourse-level devices that require contextual knowledge and cultural awareness. In this subsection, we explore the hierarchical organization of figures of speech and discuss their implications for NLP at each level, which are classified as shown in Figure 1.

Lexical Level

Figurative speech at the lexical level involves the use of language to convey meanings beyond the literal interpretation of words. Figures of speech at this level include:

- **Metaphor**: A direct comparison between two unrelated things, implying similarity (e.g., "بحر العلم").
- **Simile**: A comparison using "مثل" or "كـ" to highlight similarities (e.g., "شجاع كالأسد").

[0][1]Faculty of Sciences Ben M'Sick - Hassan II University of Casablanca, Morocco
0
Corresponding author:
Sanaa Elfilali, LTIM Laboratory, Faculty of Sciences Ben M'Sick, Hassan II University of Casablanca, Morocco
[0]Email: sanaa.elfilali@univh2c.ma

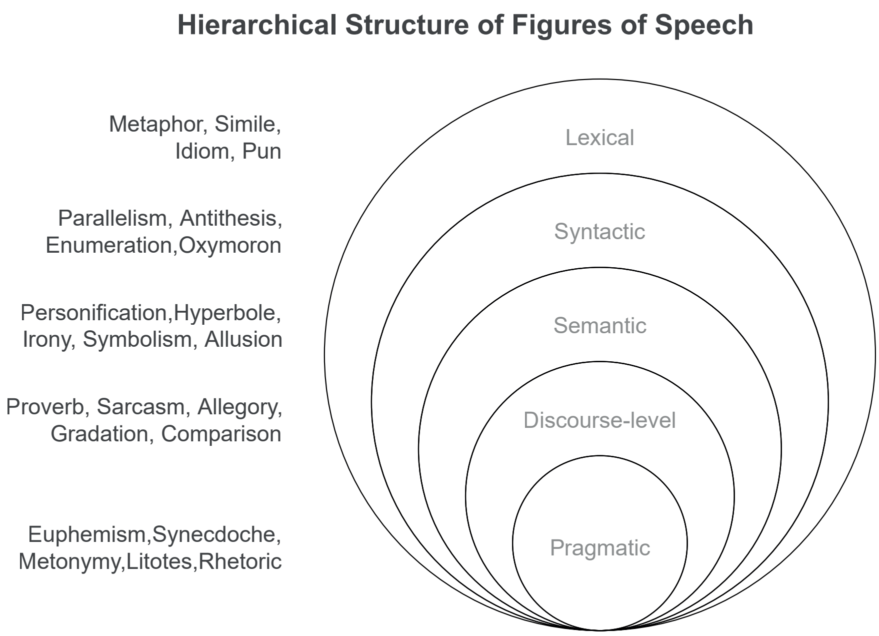

Figure 1. NLP task levels and associated figures of speech

- **Idiom**: A phrase with a figurative meaning distinct from its literal interpretation (e.g., "الطيور على أشكالها تقع").
- **Pun**: A play on words exploiting multiple meanings or similar sounds (e.g., "فتن فتىً في فتاة فتاه").
- **Oxymoron**: Juxtaposition of contradictory terms (e.g., "Deafening silence").

Metaphors play a crucial role in enriching the lexical-semantic layer of language by introducing new meanings and associations, which are essential for language evolution and adaptation [1], they are the most frequently used figurative expressions, while similes are less commonly used. Metaphors serve as a significant tool for enriching the lexical layer of language by reflecting perceptions rather than copying reality [2] [1] [3]. In addition to metaphors and similes, personifications and symbols are also used to convey deeper meanings and enhance the richness of language [4].

An oxymoron on the other hand is a figure of speech that combines contradictory terms to create a paradoxical effect. It is used to reveal a deeper truth or to create a striking impression. While oxymorons may seem self-contradictory, they can produce unexpected results when used appropriately, as they often highlight the complexity and duality of a subject [5] [6].

Figurative expressions are stored in the mental lexicon as formulaic language-use events, which help map structure to meaning in context. This phrasal/lexical memory approach suggests that all syntactic structures are encoded in the lexicon, facilitating the comprehension of both literal and figurative language [7] [8]. Recent connectionist models propose a more flexible and detailed approach to understanding word meaning, accommodating both literal and figurative language within the same framework. This model supports the interpretation of expressions based on context and rhetorical effects [8].

In summary, figurative speech at the lexical level is a complex and integral part of language that enhances communication by adding depth and richness. Understanding and effectively using figurative language can significantly improve language proficiency and comprehension.

Syntactic Level

Figurative speech at the syntactic level involves the arrangement of words and phrases to create artistic and rhetorical effects, which are used to enhance the expressiveness and impact of both spoken and written language. The structural complexity at this level poses challenges for syntactic parsing in NLP. Examples include:

- **Parallelism**: Repetition of similar structures for emphasis or rhythm (e.g., "العلم نور و الجهل عار").
- **Antithesis**: Contrasting ideas in a balanced structure (e.g., "إن الأبرار لفي نعيم، و إن الفجار لفي جحيم").
- **Enumeration**: Listing multiple items or ideas in a structured way (e.g., "Life, liberty, and the pursuit of happiness").
- **Gradation**: Progressive steps leading to a climax (e.g., "I came, I saw, I conquered").

Parallelism involves the use of similar grammatical structures in a sequence to create rhythm and emphasis. It is often used to enhance the clarity and persuasiveness of an argument by presenting ideas in a balanced and harmonious manner. This device is frequently employed in speeches and literary works to create a sense of cohesion and symmetry [9]. Antithesis is a rhetorical device that juxtaposes contrasting ideas in a parallel structure. It highlights the difference between two opposing concepts, often to emphasize a particular point or to create a dramatic effect. Examples include phrases like "I burn and I freeze" or "Her character is white as sunlight, black as midnight" [9]. This device is effective in creating a clear distinction between ideas and enhancing the persuasive power of the text [9].

Enumeration involves listing elements in a series, often to provide a comprehensive overview or to emphasize the extent of a particular point. Gradation, specifically, refers to the arrangement of words or phrases in order of increasing or decreasing importance. Both devices are used to structure information in a way that guides the reader's or listener's understanding and enhances the impact of the message [9].

Challenges and Future Directions

Understanding and distinguishing between these figurative devices can be challenging, as they often involve complex semantic and structural characteristics. Further research and educational efforts are needed to clarify these distinctions and to explore the full range of their applications in different linguistic contexts [5] [10].

Semantic Level

Figurative speech at the semantic level involves the use of language to convey meanings beyond the literal interpretation of words. This includes various figures of speech such as hyperbole, personification, allusion, symbolism, and irony, each serving unique communicative purposes and enhancing the richness of language.

Figures of speech at the semantic level involve meaning manipulation, requiring deeper understanding of word relationships and context. Examples include:

- **Personification**: Assigning human traits to non-human entities (e.g., "The wind whispered").
- **Hyperbole**: Exaggerated statements not meant to be taken literally (e.g., "I've told you a million times").
- **Irony**: Expression of meaning by using language that signifies the opposite (e.g., "What a pleasant surprise," when it's clearly not).
- **Symbolism**: Use of symbols to signify ideas or qualities (e.g., "Dove" for peace).
- **Allusion**: Indirect reference to a person, event, or work (e.g., "He has the Midas touch").

Hyperbole is characterized by exaggerated statements that are not meant to be taken literally but are used to emphasize a point or express strong emotions. It is often used in combination with other tropes like `metaphor` and `irony`, but it stands as a distinct figure of speech due to its blatant exaggeration of a scalar property to evaluate a state of affairs [11][12]. In various analyses, hyperbole is frequently identified in literary works and speeches, serving to increase emotional intensity and add emphasis [13] [14] [15].

Personification involves attributing human characteristics to non-human entities, enhancing imagery and emotional connection in language. It is commonly found in literary texts and speeches, where it serves to bring additional imagery and emotional depth to the narrative [16] [13] [14].

Allusion is a figure of speech that makes indirect references to well-known events, figures, or works, relying on the audience's familiarity with the reference to convey deeper meanings. While not extensively studied, allusion is a common rhetorical device used to enrich texts by drawing connections to broader cultural or historical contexts [17].

Similarly, symbolism involves using symbols to represent ideas or concepts, adding layers of meaning to a text. It is often used to convey complex ideas succinctly and is a powerful tool in both literature and rhetoric to evoke emotions and provoke thought [13] [15].

Irony is a figure of speech where the intended meaning is opposite to the literal meaning, often used to convey sarcasm or highlight contradictions. It frequently co-occurs with hyperbole, as ironic statements often contain exaggerated elements to enhance the ironic effect [11] [12] [18]. Irony is used to achieve various discourse goals, such as critiquing or highlighting absurdities in a situation [19] [20].

In summary, figurative speech at the semantic level enriches communication by allowing speakers and writers to convey complex ideas, emotions, and evaluations in a nuanced and impactful manner. Each figure of speech serves distinct purposes, contributing to the overall effectiveness and depth of language.

NLP must handle ambiguity and cultural knowledge to process these effectively.

Discourse-Level

At the discourse level, figures of speech span multiple sentences or larger text units, making coherence and thematic analysis crucial. Examples include:

- **Proverb**: A short saying with general truth or advice (e.g., "Actions speak louder than words").
- **Sarcasm**: Use of irony to mock or convey contempt (e.g., "Oh, great, another homework assignment").
- **Allegory**: An allegory is a satirical figure that uses a symbolic image or metaphor to express a critical thought or idea. Allegories are often used to covertly criticize a political system, social problem, or social phenomenon. [21] (e.g., "Animal Farm" as an allegory for totalitarianism).
- **Satire**: Satire is the use of humor, irony, exaggeration, or ridicule to expose and criticize flaws in society, politics, or individuals.

Figurative speech at the discourse level, including allegory, sarcasm, comparison, and proverbs, plays a significant role in communication by enhancing meaning, conveying complex ideas, and influencing audience perception. This synthesis explores the use and impact of these forms of figurative language in various contexts.

Figurative language serves multiple discourse goals, such as comparing similarities, adding emphasis, and being humorous. These goals vary depending on the context and modality of communication. For instance, similes and metaphors are often used to highlight similarities, while irony and sarcasm can be employed to counter or critique preceding statements in public discourse [19] [22] [23]. In political discourse, figurative language can amplify or soften content, saving linguistic means while also serving emotional and attractive functions [24].

Irony and sarcasm are prevalent in public and political discourse, often used to challenge or undermine previous statements. They require the retention of the original metaphorical expression as a reference point to achieve the intended effect. This interplay keeps the metaphorical meaning alive and relevant in ongoing discussions [23]. These forms of figurative speech are also crucial in shaping public opinion and influencing audience perception in media-political communications [24].

Proverbs utilize metaphorical language to convey wisdom and cultural values. They reflect the worldview of a community and are used to communicate moral lessons and social norms. The cognitive-analogical processes in proverbs create mental images that reflect cultural peculiarities, such as attitudes towards life and morality [25] [26]. In academic contexts, proverbs can serve as guiding principles, offering insights and experiences relevant to personal and educational journeys [26].

Pragmatic Level

The pragmatic level focuses on figures of speech that rely on context, speaker intent, and shared knowledge. These are the most challenging for NLP due to their reliance on external factors. Examples include:

- **Euphemism**: Mild or indirect terms for something harsh or unpleasant (e.g., "Passed away" for "died").
- **Synecdoche**: Part used to represent the whole, or vice versa (e.g., "Wheels" for a car).
- **Metonymy**: Use of a related term to represent something (e.g., "The crown" for monarchy).
- **Litotes**: Understatement for emphasis (e.g., "Not bad" for "good").
- **Rhetoric**: Artful use of language to persuade or impress (e.g., "Ask not what your country can do for you").

Figurative speech, encompassing devices such as euphemism, litotes, synecdoche, metonymy, and rhetoric, plays a significant role in pragmatics, which is the study of language in context. These figures of speech are not merely decorative elements but are integral to effective communication, influencing how messages are conveyed and interpreted.

Euphemism is a rhetorical technique that replaces taboo or harsh terms with more polite or less direct expressions. It functions as an indirect speech act, often violating the Cooperative Principle to adhere to the Politeness Principle and Face Theory, thus maintaining social harmony and personal relationships [27] [28]. Euphemisms are context-dependent and serve purposes such as politeness, camouflage, and beautification [28].

Litotes, on the other hand, involves understatement by using double negatives or a negative to affirm a positive. It is a pragmatic tool that can either downplay a statement or, paradoxically, emphasize it by awakening the idea of more [29]. The pragmatic divergence between euphemism and litotes lies in their communicative intent—euphemism softens the impact, while litotes can subtly highlight it [29].

Synecdoche and metonymy are figures of speech that condense complex ideas into simpler terms. Synecdoche involves using a part to represent the whole or vice versa, while metonymy uses a related concept to stand in for another. Both draw rhetorical energy from a cluster of meanings, preserving dominant connotations and facilitating efficient communication [30]. These devices are crucial in rhetoric as they encapsulate broader narratives or ideas into memorable expressions [30].

The use of rhetoric, including euphemism, metaphor, and metonymy, often breaches conversational principles, such as the Cooperative Principle, to achieve specific communicative goals [31]. This breach is not a failure of communication but a strategic move to convey nuanced meanings, manage politeness, or influence the listener's perception [31].

Research Methodology

This section outlines the systematic approach adopted for conducting the literature review, ensuring a comprehensive and unbiased synthesis of relevant research related to figurative speech detection. In the following subsections we describe the steps conducted in the present research synthesis

Research Questions

The literature review was guided by the following research questions:

1. What datasets and resources are commonly utilized for this task?
2. What are the predominant computational methods used for figurative speech detection?
3. How do existing methods perform across different types of figurative language ?
4. Which languages have seen particular advancements in this field ?

Search strategy

The process adheres to the PRISMA *(Preferred Reporting Items for Systematic Reviews and Meta-Analyses)* framework, which provides a structured methodology for identifying, screening, and including relevant studies. A visual representation of this process is provided in Figure 2.

Articles were queried from both Scopus and Google Scholar, using the research tool *Publish or Perish*. The search was conducted by specifying keywords that align with the research questions outlined earlier. We used the following template to query research articles for each of our 24 targeted figures of speech:

```
("<FIGURE_ALIAS_1>" OR "<FIGURE_ALIAS_2>"
OR ...)  AND ("figurative language" OR
"text analysis" OR "stylistic analysis")
AND ("identification" OR "detection"
OR "classification")
```

Search results were exported and compiled into a CSV file for further processing. The total number of retrieved records was 6331, of which 4386 were removed upon preliminary filtering. The remaining 1945 records underwent screening based on titles, keywords and abstracts to assess their relevance to the inclusion criteria.

Next, the full-text articles of 95 potentially relevant records were reviewed for eligibility. Studies were excluded if they did not meet the criteria described in section . The subsequent number of included studies was 59, as summarized in the PRISMA diagram (Figure 2). These

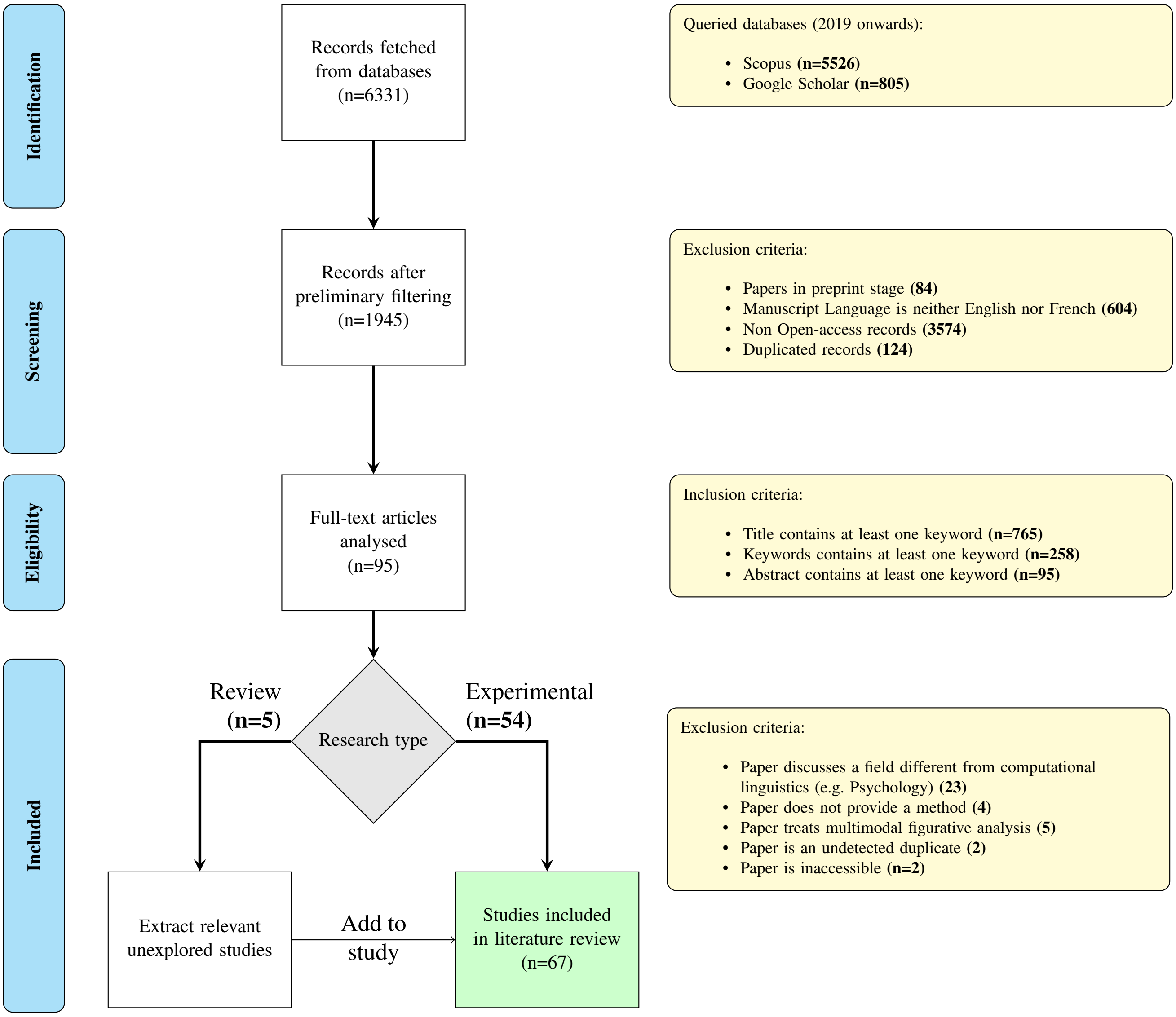

Figure 2. PRISMA flow diagram illustrating the steps followed in the study selection process.

studies were identified through the filtering process, which excluded duplicates and studies that did not meet the inclusion criteria.

Since we are interested in experimental results, the focus of this review was primarily on 54 studies classified as experimental. A smaller subset of 5 review studies was included to identify unexplored gaps and extract relevant insights that could complement the experimental findings. Through this process, additional relevant papers were identified and added to the analysis, bringing the final number of included studies to 67. This approach ensures a comprehensive synthesis, combining experimental data with broader insights from review studies to address the research questions effectively.

Inclusion and exclusion criteria

The selection process for studies was conducted in multiple stages, with inclusion and exclusion criteria applied at each level of the pipeline. Below, we detail the criteria applied at each stage:

Initial Search The initial search retrieved records 5826 from Scopus and 805 from Google Scholar, spanning publications from 2019 onward. The following exclusion criteria were applied at this stage:

- Papers in the preprint stage (n=84).
- Manuscripts in a language other than English or French (n=604).
- Non-open-access records (n=3574).
- Duplicated records removed after deduplication (n=124).

Keyword Filtering To narrow down the results, inclusion criteria were applied based on the presence of specific keywords:

- The title contains at least one of the specified keywords, resulting in 765 papers.
- The keywords section contains at least one of the specified keywords, resulting in 258 papers.
- The abstract contains at least one of the specified keywords, resulting in 95 papers.

Full-Text Screening The remaining studies were subjected to full-text screening to ensure relevance and methodological rigor. Exclusion criteria at this stage included:

- Papers discussing fields unrelated to computational linguistics (e.g., psychology) (n=23).
- Papers that do not provide a method (n=4).
- Papers treating multimodal figurative analysis (n=5).
- Papers marked as undetected duplicates (n=2).
- Papers that were inaccessible (n=2).

By systematically applying these criteria at different stages, the final pool of studies represents a highly curated dataset, ensuring relevance and quality for addressing the research objectives.

Data analysis

Table 11 provides an overview of the **67** literature analyzed in this study, categorizing works based on their publication type, including journals and conference proceedings. A total of **42** journal articles, **24** conference papers and **1** thesis were reviewed, covering a range of figurative speech detection related research. The selected works span publication years from **2019** to **2024**, with a few selected works dating from previous years, highlighting the evolution of research in this domain. The diversity of venues and publication years underscores the growing academic interest and advancements in applying various computational approaches to detect figurative language in different contexts.

Related Work

Figurative speech detection has gained significant attention in recent years due to its crucial role in natural language understanding. The inherent complexity of figurative language, such as metaphor, irony, and sarcasm, presents unique challenges for computational approaches, primarily due to their ambiguity and cultural dependencies. Several literature reviews have previously explored different aspects of figurative speech detection; however, our work provides a more comprehensive synthesis across multiple levels of figurative style comprehension and languages, addressing existing gaps and offering new insights.

Existing Literature Reviews

Prior studies have attempted to summarize advancements in figurative speech detection. Table 12 presents a summary of the key surveys in this domain.

These studies have focused on various aspects such as metaphor detection in English, irony detection in social media, and multilingual challenges. However, they often exhibit limitations such as a lack of focus on low-resource languages, or an emphasis on specific figurative styles. Our review aims to bridge these gaps by offering a comprehensive analysis of the field, providing a computational perspective that categorizes figurative speech based on its level of linguistic processing within the NLP ecosystem—ranging from lexical-level figures to those requiring deeper semantic and pragmatic understanding.

Methodologies for Figurative Speech Detection

Over the years, several computational methods have been employed for figurative speech detection, which can be categorized into four major approaches:

Lexicon-Based Approaches Lexicon-based methods rely on manually or automatically curated lexicons such as WordNet and ConceptNet to detect figurative expressions. These approaches leverage linguistic resources to capture meaning and context, making them interpretable but often lacking generalizability, nevertheess, the remain a potential alternative for languages with low resources. Studies such as [32, 33, 34] have demonstrated their effectiveness in detecting oxymorons, idioms and metaphors, respectively for Bengali, Gujarati and Ukrainian languages.

Machine Learning-Based Approaches Traditional machine learning methods, including Support Vector Machines (SVM), Naïve Bayes, and Decision Trees, have been widely adopted for figurative speech detection. Feature engineering techniques such as TF-IDF, n-grams, and word embeddings (e.g., Word2Vec, GloVe) have played a crucial role in enhancing the performance of these methods [35, 36, 37]. However, these methods often struggle with domain adaptation and require extensive labeled data. Thus, novel feature engineering techniques were combined with ML models to improve performance, including syntactic and semantic features such as dependency parsing, sentiment scores, and context-aware embeddings. Studies have shown that incorporating such features allows models to capture deeper linguistic patterns, improving generalizability across different figurative speech styles and domains [38, 39, 40].

Deep Learning Approaches Recent advancements in deep learning have revolutionized figurative speech detection with models such as LSTMs, CNNs, and transformer-based architectures like BERT and RoBERTa, whether for feature extraction [41] or fine-tuning [42]. These models leverage contextual embeddings to achieve state-of-the-art performance, particularly in high-resource languages. Transformer-based architectures have have especially been employed in higher levels of figurative speech detection, namely semantic and discourse-level (e.g. sarcasm, satire, and irony), due to their ability to capture larger contextual information than other recurrent models such as LSTMs and GRUs [43, 44]

Hybrid Approaches Hybrid methods combine rule-based and predictive approaches to leverage the strengths of both. Studies such as [45, 46] have shown that hybrid approaches can improve figurative speech detection accuracy by incorporating linguistic features alongside feature-extraction models. These methods typically utilize rule-based techniques to capture explicit figurative expressions and domain-specific knowledge, while machine learning models generalize patterns from data to handle implicit and context-dependent cases. For instance, lexicon-derived sentiment scores and part-of-speech (POS) tags are often combined with statistical learning models such as SVM and random forests to enhance classification performance. [46] Additionally, deep learning architectures have been augmented with lexicon features to provide interpretability and mitigate the data scarcity issues prevalent in low-resource languages. Despite their advantages, hybrid methods face challenges such as increased computational complexity and the need for extensive feature engineering to achieve optimal integration of rule-based and statistical techniques. Future research in this area is exploring the potential of transformer-based models to seamlessly incorporate linguistic rules while maintaining high efficiency and scalability.

Existing Figurative Speech Datasets

Several annotated datasets have been developed for figurative speech detection, with a primary focus on high-resource languages, particularly English. Early datasets, such as TroFi [47] and SemEval [48], introduced benchmark corpora for metaphor and irony detection. However, reviews have highlighted critical challenges, including class imbalance and a lack of diversity in figurative styles and languages. A more detailed analysis of dataset characteristics is presented in section .

Analysis

This section provides an overview of trends and patterns observed across the studies, setting the stage for the level-specific analyses. Each research question (RQ1–RQ4) is addressed in the subsequent sections.

RQ1: What datasets and resources are commonly utilized for this task?

Analysis of Dataset Utilization Across Figurative Speech Levels Table 1 presents an overview of the distribution of dataset sources utilized across different figurative speech levels, as reported in the articles analyzed in this study. Each number in the table indicates how many studies employed a particular dataset type for a specific figurative speech style.

According to Table 1, the discourse-level style is the most frequently studied, with **26** mentions across various dataset sources. This trend can be attributed to the inclusion of **sarcasm**, a prevalent figurative speech style that often manifests in extended textual contexts, such as social media conversations, news articles, and dialogues. The dominance of **Social Media datasets** (**12 mentions**) in discourse-level tasks further highlights the significance of user-generated content in capturing the nuances of sarcasm, which frequently relies on contextual cues and conversational exchanges to convey meaning. The reliance on discourse-level sources suggests a research preference for analyzing sarcastic expressions in naturally occurring interactions rather than structured or formal texts.

Research benchmarks are the most commonly cited dataset sources, appearing in **23 studies** across multiple figurative speech styles. Their widespread adoption suggests a preference for structured and well-annotated datasets, which are critical for benchmarking machine learning models. The highest concentration is observed in the semantic style (**8 mentions**), followed by discourse-level (**7 mentions**), emphasizing the complexity and need for reliable data in these areas.

Literary works, including religious texts such as Qur'an, are predominantly used in **lexical** and **pragmatic** studies (**4 mentions each**), reflecting their suitability for analyzing rich, contextually dense figurative expressions found in books and classic texts. Their limited presence in **discourse-level** studies can be attributed to the nature of the figurative styles analyzed within this category—namely sarcasm, satire, and, to a lesser extent, proverbs—which are more commonly encountered in contemporary and conversational sources such as social media and news media. This suggests that the observed distribution aligns with the natural occurrence of these styles rather than indicating an underutilization of literary sources.

Despite the growing availability of online text resources, web-based content and online reviews were only referenced in **2 studies each**. This suggests a potential underutilization of informal, real-world data sources in figurative speech detection research. Given the richness and diversity of online content, future work could benefit from leveraging these sources to capture more spontaneous and diverse figurative expressions.

The syntactic style level has the lowest overall study count (**6**), highlighting a significant research gap in this area. Figurative speech styles that require syntactic analysis, such as antithesis and chiasmus, often rely on grammatical structures that demand specialized datasets with detailed syntactic annotations. The limited availability of such resources may hinder advancements in this domain. Notably, research benchmarks and news media are the primary sources used in syntactic analysis, but their infrequent usage indicates a need for more comprehensive syntactically-annotated corpora to support deeper exploration of figurative language at this level. The high reliance on research benchmarks and social media data indicates a potential overfitting of research efforts towards these domains, potentially limiting the generalizability of findings. The lower number of studies focusing on syntactic and pragmatic styles suggests a need for more diverse and representative datasets that capture these linguistic aspects. Encouraging the exploration of underutilized sources, such as web-based content and online reviews, may enhance the robustness of figurative speech detection methods.

Overview of Language Coverage Table 2 provides an overview of the datasets used in figurative speech detection studies across various languages. The data highlights significant disparities in language representation, with a clear dominance of English datasets, while other languages, such as Bengali and Gujarati, remain underrepresented.

English is the most extensively covered language, with **41 studies** utilizing datasets sourced from diverse domains, including *social media* (e.g., Twitter, Reddit), *research benchmarks* (e.g., GWN, HYPO, MOH2015), and *literary works*. This dominance reflects the widespread availability of English datasets and the strong focus of research on English figurative speech detection.

Limited Resources for Low-Resource Languages In contrast, languages such as **Gujarati** and **Bengali** have received minimal attention, with only a few studies relying on datasets such as *GujaratiLexicon.com* and *TDIL*. This indicates a clear research gap, as the lack of annotated resources for these languages hinders the development of figurative speech detection models beyond high-resource languages.

Arabic and French: Emerging Research Interest Arabic (**7 studies**) and French (**2 studies**) exhibit moderate coverage, with datasets sourced from *social media* and *religious texts* such as the *Quran*. Despite their linguistic

Table 1. Distribution of dataset utilization across figurative speech levels.

Style Level	Lexical	Syntactic	Semantic	Discourse-level	Pragmatic	All
Broadcast Media	0	1	0	1	0	2
Kaggle	1	0	0	1	0	2
Literary Works	4	2	1	0	4	8
News Media	0	1	1	3	1	6
Online Reviews	1	0	1	0	0	2
Research Benchmark	5	2	8	7	4	23
Shared Tasks	1	0	5	3	2	10
Social Media	0	0	7	12	1	19
Web-based Content	0	0	0	1	1	2
Total	11	6	18	26	12	67

richness and cultural significance, research efforts in these languages remain limited compared to English.

Diversity of Data Sources Across Languages While English datasets originate from a variety of sources, other languages show a narrower range of sources. For example:

- **Arabic** datasets primarily stem from religious texts and social media.
- **German** datasets are limited to *German dramas* and a few benchmark datasets.
- **Spanish** datasets are mostly derived from shared tasks and research benchmarks.

Shared Tasks and Benchmark Resources The presence of datasets from shared tasks, such as *SemEval*, demonstrates efforts to standardize evaluation across multiple languages, particularly in English and Spanish. However, other languages see limited participation in these initiatives, highlighting the need for multilingual benchmark efforts.

Web-Based Content as an Untapped Resource Despite the wealth of available online content, web-based datasets (e.g., Wikipedia, phrases.co.uk) are sparsely used across all languages, with only **2 studies** utilizing them. This suggests an opportunity for future research to harness online content as a valuable resource for figurative speech detection.

Key Research Gaps and Future Directions The findings from this table underscore several critical gaps:

1. **Expanding Multilingual Coverage:** There is a need for more datasets in underrepresented languages such as Bengali and Gujarati.
2. **Diversifying Dataset Sources:** Researchers should explore alternative sources such as online reviews and broadcast media across all languages.
3. **Cross-Linguistic Comparisons:** A broader range of language-specific datasets could enable better cross-linguistic comparisons in figurative speech analysis.

Granularity Trends in Figurative Speech Styles As can be seen in Table 3, lexical figurative style detection is mostly at the phrase level because figurative expressions, such as idioms and metaphors, often manifest as fixed or semi-fixed phrases rather than isolated words. These expressions derive their meaning from the collective interpretation of multiple words rather than individual lexical items. For instance, detecting an idiom like خياران أحلاهما مر requires analyzing the phrase as a whole rather than the separate meanings of individual words. Phrase-level analysis allows models to capture contextual and syntactic relationships within short text spans, making it more effective for lexical figurative language, where meaning heavily depends on word combinations and their conventionalized usage.

This pattern is more emphasized in syntactic figurative style detection is due to the inherent structural complexity of syntactic figurative language, such as syntactic enumerations, parallelisms, and word order inversions (chiasmi). These phenomena often span multiple words and depend heavily on phrase-level syntactic structures, such as noun phrases, verb phrases, or prepositional phrases, to convey their figurative meaning. Unlike lexical figurative language, which primarily relies on fixed expressions, syntactic figurative styles often involve variations in word order, dependency relations, and phrase embeddings within sentences. Thus, focusing on phrases allows researchers to capture the hierarchical and dependency-based structures that are crucial for syntactic analysis. Additionally, syntactic figurative styles often require deeper parsing techniques, such as constituency or dependency parsing, which naturally operate at the phrase level rather than individual words or entire sentences.

Additionally, we observe that semantic and discourse-level figurative speech styles have a high concentration of annotation at the sentence level (**84.21%** and **85.19%**, respectively). This pattern arises because semantic-level figurative language often requires understanding the full context of a sentence to capture meaning shifts, such as hyperboles and personification, which rely on broader contextual cues rather than isolated words or phrases. Many figurative expressions at the semantic level depend on implicit relationships and word sense disambiguation, which are better analyzed when the full sentence context is available. Discourse-level figurative language, such as sarcasm, irony, and satire, typically spans entire sentences or even multiple sentences. The pragmatic cues, tone, and contextual dependencies needed to understand these styles often emerge across the entire sentence rather than individual words or phrases. Sentences provide a more complete unit for understanding dialogue dynamics and speaker intent, which are crucial for discourse-level analysis.

The distribution of annotation granularity for pragmatic figurative speech exhibits a notable spread across multiple levels, reflecting the diverse linguistic structures these expressions encompass. Unlike other figurative speech categories that exhibit a clear preference for specific

Table 2. Datasets used across covered languages

Data Source	Arabic	Bengali	English	French	German	Gujarati	Low-resource languages	Spanish
Broadcast Media			BBC Radio 4 Show				TV Comedy shows	
Kaggle			Kaggle					
Religious texts Literary Works	Quran		Books *The catcher in the Rye*		German dramas by Friedrich Schiller	Books	Literary works Amharic idiom books	
News Media			NYT Annotated Corpus News headlines	News websites			Newspaper	
Online Reviews			Amazon Fine Foods Reviews Product reviews					
Research Benchmark	Other research	TDIL	Barbieri'15 GWN HYPO Moh2015 Other research Ptacek2014 RELOCAR Reyes2013 Riloff TOEFL (AllPOS) TOEFL (Verbs) TroFi VUA (AllPOS) VUA (VERBS) WIMCOR WIMCOR (Tokens)	Other research			SaRoCo	Barbieri'15 IroSVA Salas'17
Shared Tasks	IDAT shared task	FIRE	CONLL EMNLP 2022 Irony SemEval2018 MUStARD SEMEVALLOC SemEval SemEval2018					
Social Media	Twitter	Facebook	Reddit SARC 2.0 Twitter				Sarc-H Twitter	Twitter (Shared task)
Web-based Content			Wikipedia phrases.co.uk			Websites		

Table 3. Granularity distribution across figurative speech styles (values represent percentages within each style level, with the total column representing absolute counts).

Style Level	Token	Phrase	Clause	Sentence	Discourse	Document	Total
Lexical	27.27	54.55	9.09	9.09	0.00	0.00	11
Syntactic	16.67	66.67	16.67	0.00	0.00	0.00	6
Semantic	0.00	0.00	5.26	84.21	5.26	5.26	18
Discourse-level	0.00	0.00	0.00	85.19	7.41	7.41	26
Pragmatic	23.08	15.38	7.69	53.85	0.00	0.00	12

granularity levels, pragmatic expressions are annotated at varying levels, from token to sentence. This distribution suggests that pragmatic expressions often require flexible annotation strategies, as their meaning can be conveyed effectively at different levels of textual context. The predominance of sentence-level annotations indicates that pragmatic expressions frequently rely on the broader syntactic and semantic context for accurate interpretation. Additionally, the presence of annotations at both token and phrase levels highlights the adaptability of these expressions in shorter textual spans, as in identifying metonymies within a text, further reinforcing the need for a multi-level approach in their analysis. Table 4 shows the proportions of labeling scopes for each level of figurative speech.

Table 4. Annotation schemes (%) used across different figurative speech styles.

Style Level	Binary	IOB	Multiclass
Lexical	63.64	9.09	27.27
Syntactic	50.00	16.67	33.33
Semantic	94.44	0.00	5.56
Discourse-level	88.46	0.00	11.54
Pragmatic	33.33	8.33	58.33

The distribution of annotation schemes across different figurative speech styles reveals notable patterns that align with the complexity and structural characteristics of each style. Binary annotation dominates the majority of styles, particularly in discourse-level and semantic categories, where 88.46% and 94.44% of the datasets employ binary labels, respectively. This suggests a tendency to simplify the classification of these styles, due to the possibility of overlapping multiple figures within the same span. In contrast, the lexical and syntactic styles exhibit a higher degree of multiclass annotation, with 27.27% and 33.33%, respectively, reflecting the inherent structural complexity of these styles, which may require more detailed categorization. The relatively low adoption of the IOB scheme, particularly in discourse-level and semantic styles, stems from the nature of these figures of speech that require a full understanding of the context more than the identification of key phrases. The pragmatic style stands out with a balanced distribution of annotation schemes, with multiclass annotation reaching 58.33%, underscoring the nuanced nature of pragmatic expressions that often require multi-faceted labeling approaches.

Dataset Availability across Style Levels Table 5 presents the availability of datasets across five figurative speech style levels—Lexical, Syntactic, Semantic, Discourse, and Pragmatic—divided into benchmark and non-benchmark categories. Each row within the table represents a specific availability type, including *Available upon request*, *Creative Commons License*, *Publicly available*, *Unavailable*, and *Unspecified*.

Table 5 presents the availability of datasets across different figurative speech style levels, segmented into benchmark and non-benchmark categories. The data reveals several notable trends.

First, publicly available datasets are disproportionately concentrated in the *Semantic* and *Discourse-level* styles, with 100% of syntactic benchmark datasets and 50% of non-benchmark discourse-level datasets being publicly accessible. This suggests that these styles benefit from a higher degree of open access, likely due to their relevance to widely shared tasks, such as sentiment analysis and sarcasm detection.

Conversely, *Lexical* benchmark datasets exhibit a high percentage of unavailability (60%), reflecting potential challenges in releasing such datasets due to licensing constraints or reliance on proprietary sources, such as published books or annotated texts. This pattern aligns with the intrinsic complexity of lexical figurative speech, which often requires substantial manual annotation efforts and relies on diverse, fragmented sources.

The *Pragmatic* style level presents a more balanced distribution, with both publicly available and unavailable datasets, indicating that while some resources are shared, others remain restricted. This might be attributed to the diverse applications of pragmatic figurative speech in specific domains, such as metonymy resolution or rhetorical question analysis, where domain-specific datasets are less frequently shared.

Unspecified availability rates are particularly concerning at the *Semantic* level, where 21.43% of datasets do not clearly state their access conditions. This lack of clarity hinders reproducibility and limits the practical applicability of research findings in this area.

RQ2: What are the predominant computational methods used for figurative speech detection?

Common feature extraction techniques Figure 3 presents an overview of the most commonly used feature extraction techniques for figurative speech detection, highlighting the relative prevalence of different feature types.

The figure reveals that **embedding-based techniques** are the most commonly used method, accounting for **21.87%** of the total. This dominance indicates the widespread reliance

Table 5. Availability percentage of benchmark and non-benchmark datasets across figurative speech style levels.

Data Source Type	Data Availability	Lexical	Syntactic	Semantic	Discourse	Pragmatic
Benchmark	Available upon request	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
	Creative Commons License	0.00 %	0.00 %	11.11 %	0.00 %	0.00 %
	Publicly available	20.00 %	100.00 %	22.22 %	25.00 %	75.00 %
	Unavailable	60.00 %	0.00 %	44.44 %	75.00 %	25.00 %
	Unspecified	20.00 %	0.00 %	22.22 %	0.00 %	0.00 %
Not benchmark	Available upon request	0.00 %	0.00 %	0.00 %	5.00 %	0.00 %
	Creative Commons License	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
	Publicly available	28.57 %	75.00 %	50.00 %	40.00 %	33.33 %
	Unavailable	57.14 %	0.00 %	28.57 %	50.00 %	55.56 %
	Unspecified	14.29 %	25.00 %	21.43 %	5.00 %	11.11 %

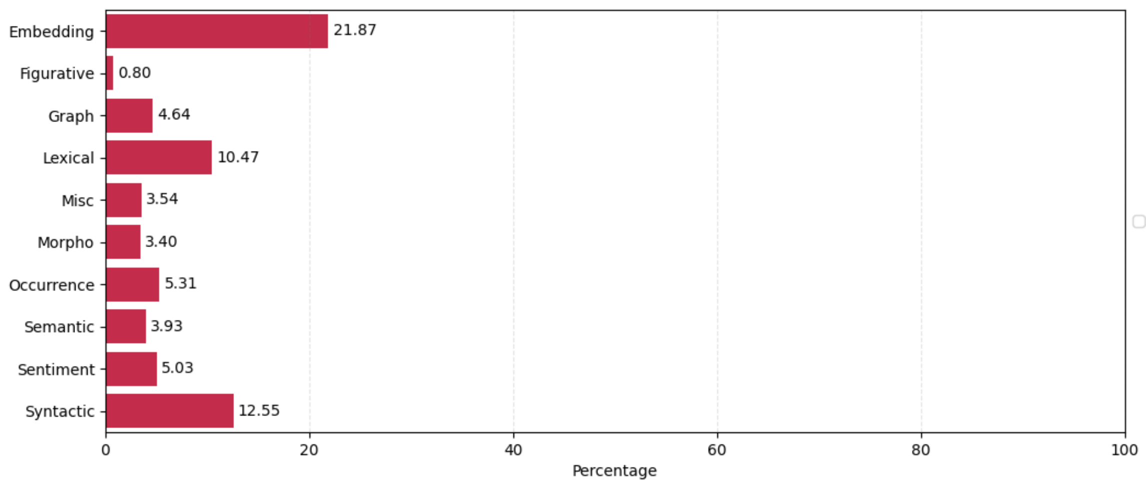

Figure 3. Distribution of computational approaches used in figurative speech detection studies.

on pre-trained language models and distributed representations to capture contextual and semantic nuances of figurative language. Embeddings provide a rich representation of words and phrases, making them particularly effective for identifying figurative expressions such as metaphors and idioms.

Syntactic features, constituting **12.55%**, are the second most used, emphasizing their importance in detecting structural patterns that characterize various figurative speech styles, such as chiasmus and parallelism. These features are often combined with rule-based approaches to identify specific syntactic constructs indicative of figurative expressions.

Lexical features, accounting for **10.47%**, continue to be relevant, relying on handcrafted features such as word lists, dictionaries, and frequency-based statistics. Lexical analysis helps identify common patterns and known figurative phrases, making it a useful feature set, especially in conjunction with rule-based systems.

Sentiment features, contributing **5.03%**, play a notable role in figurative speech detection, particularly for detecting sarcasm, irony, and hyperbole, where emotional tone and intensity serve as key indicators.

Occurrence-based features (5.31%) represent statistical properties such as term frequency and word distribution, which assist in detecting patterns specific to figurative expressions by measuring their regularity or unexpectedness within text.

Graph-based (4.64%), miscellaneous (3.54%), morphological (3.40%), and semantic (3.93%) features contribute smaller portions, indicating that while these methods are useful, they are often supplementary to embeddings, syntactic, and lexical approaches.

Figurative-specific features, representing only **0.80%**, suggest that explicitly designed features tailored for figurative language detection (e.g., humor markers, exaggeration cues) are relatively underexplored compared to general-purpose linguistic features.

The analysis suggests that modern figurative speech detection heavily relies on **embedding-based techniques** due to their ability to capture deep contextual meaning, followed by **syntactic and lexical features** that provide interpretable rule-based insights. Sentiment and occurrence-based features also play a supporting role, helping refine the detection process by offering additional context and statistical cues. The relatively low representation of figurative-specific features indicates an opportunity for further research into domain-specific feature engineering tailored to figurative language detection.

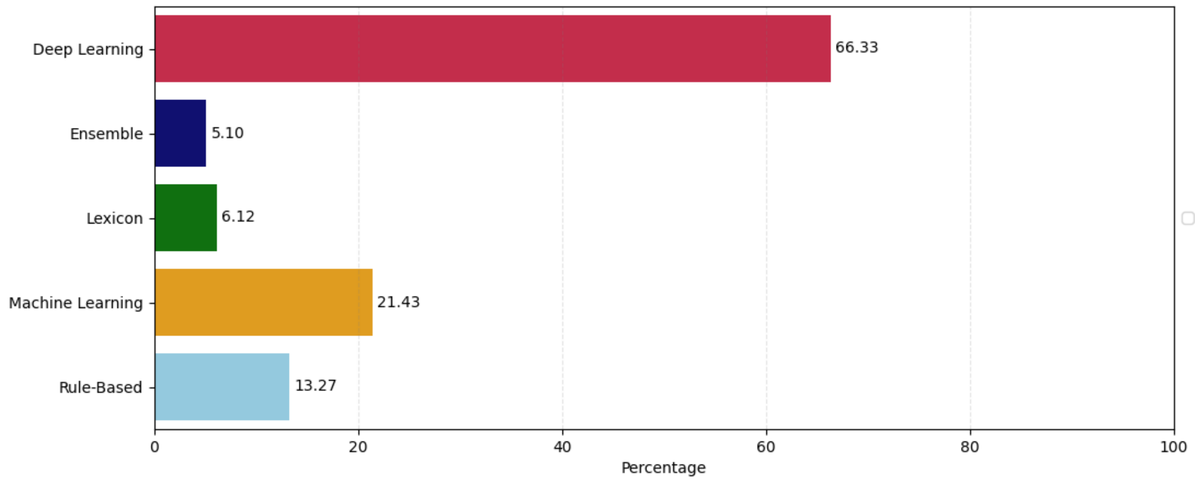

Figure 4. Distribution of methodological approaches used in figurative speech detection studies.

Commonly Used Models Figure 4 illustrates the distribution of different model categories employed in figurative speech detection, highlighting the dominance of deep learning-based approaches.

The figure reveals that **deep learning models** are the most widely used, accounting for **66.33%** of the total. This overwhelming preference for deep learning techniques indicates the increasing reliance on complex neural networks, such as transformers, convolutional neural networks (CNNs), and recurrent neural networks (RNNs), which are well-suited for capturing contextual and sequential dependencies in figurative language.

Machine learning models, contributing **21.43%**, remain an essential part of figurative speech detection, with traditional classifiers such as Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression offering effective solutions for structured and feature-based analysis. These models are particularly useful in scenarios where interpretability and explainability are important.

Rule-based approaches, comprising **13.27%**, continue to play a role in figurative speech detection, especially for well-defined patterns and fixed expressions. Techniques such as regular expressions and syntactic rules are often employed for figures of speech that follow predictable structures, such as oxymorons and chiasmi.

Lexicon-based methods, accounting for **6.12%**, are commonly used to leverage predefined dictionaries and resources such as sentiment lexicons and concreteness databases. These approaches are particularly valuable for detecting figurative expressions that rely on word meanings and associations, such as metaphors and idioms.

Ensemble models, though representing a smaller share at **5.10%**, demonstrate the growing interest in combining multiple modeling approaches to improve performance. By leveraging a combination of deep learning, machine learning, and rule-based techniques, ensemble methods aim to achieve higher robustness and accuracy.

The analysis suggests that deep learning models are the most commonly used due to their superior capability in handling complex, context-rich data, followed by machine learning models that provide a balance between performance and interpretability. While rule-based and lexicon-based methods remain relevant, their usage is significantly lower compared to data-driven approaches. The presence of ensemble methods highlights the potential benefits of combining various techniques to enhance figurative speech detection performance.

Achieved performance Table 6 presents the statistical analysis of various performance metrics utilized across figurative speech detection methods. The most commonly reported metric is the F1-Score, appearing in 78 experiments, with an average of 80.61% and a standard deviation of 12.52, indicating consistent model performance across different studies. Accuracy is the second most frequently reported metric (n=54), with a mean of 85.83%, highlighting its widespread use in evaluating classifier effectiveness.

Less frequently reported metrics include AUC (n=5) and Loss (n=2), indicating that while precision-based measures are predominant, optimization-based and ranking-based metrics are underutilized. Additionally, the presence of unreported or unspecified metrics (n=11) suggests variability in evaluation practices across studies.

Overall, the high variance in precision and recall metrics (Std = 14.53 and 16.54, respectively) points to the challenges in consistently detecting figurative language across diverse datasets. The use of median values (e.g., 93% for Accuracy, 90% for Specificity) suggests that most models achieve high performance, with some outliers lowering the overall mean values.

To ensure a comprehensive evaluation of figurative speech detection methods, future research should emphasize reporting consistency and include additional evaluation measures such as Kappa and Jaccard Score, which currently have limited representation in the literature.

Usage Trends Figure 5 illustrates the trends in model usage over time, showing a clear dominance of **predictive approaches**, which have seen a sharp increase since 2018, peaking in 2020 and maintaining a strong presence in subsequent years. This growth can be attributed to advancements

Table 6. Descriptive statistics of performance metrics used in figurative speech detection. The table presents the frequency (Count), mean, standard deviation (Std), and distribution percentiles (Min, 25%, 50%, 75%, Max) for each reported metric, highlighting the prevalence of F1-Score and Accuracy as the most commonly used evaluation measures.

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
AUC	5	88.0000	8.6891	74.0000	85.0000	93.0000	94.0000	94.0000
Accuracy	54	85.8333	9.8876	58.0000	79.2500	89.5000	93.0000	100.0000
Avg. Precision	1	23.0000	-	23.0000	23.0000	23.0000	23.0000	23.0000
F1-Score	78	80.6154	12.5255	44.0000	71.0000	81.5000	90.0000	100.0000
Hamming Loss	1	0.0300	-	0.0300	0.0300	0.0300	0.0300	0.0300
Jaccard Score	1	92.0000	-	92.0000	92.0000	92.0000	92.0000	92.0000
Kappa	1	98.0000	-	98.0000	98.0000	98.0000	98.0000	98.0000
Loss	2	0.6305	0.8761	0.0110	0.3208	0.6305	0.9402	1.2500
Precision	55	77.9455	14.5289	41.0000	67.5000	82.0000	87.0000	100.0000
Recall	55	77.5636	16.5381	38.0000	66.5000	83.0000	90.5000	100.0000
Specificity	1	90.0000	-	90.0000	90.0000	90.0000	90.0000	90.0000
Unspecified	11	-	-	-	-	-	-	-

in deep learning techniques such as transformers and contextual embeddings, which have demonstrated superior capabilities in capturing complex figurative language patterns. In contrast, **rule-based approaches**, which were previously popular, exhibit relatively stable usage over time but have been overshadowed by data-driven methods due to their limited scalability and adaptability to diverse datasets.

Hybrid models, which combine predictive and rule-based techniques, have shown sporadic growth, with notable spikes around 2020, indicating their potential in leveraging the strengths of both approaches. However, their adoption remains relatively lower compared to predictive methods, suggesting that their complexity and integration challenges may limit widespread application.

The performance analysis suggests that while current models achieve high accuracy and balanced precision-recall trade-offs, there is room for improvement, particularly in recall. The trends highlight a clear shift toward predictive methods, driven by advancements in deep learning, while rule-based and hybrid approaches continue to play supportive roles in addressing specific figurative speech detection challenges.

RQ3: How do existing methods perform across different types of figurative language ?

Key Performance Metrics in Figurative Speech Detection Figure 6 illustrates the distribution of evaluation metrics used to assess figurative speech detection methods across different style levels. The analysis reveals that the most commonly utilized metric is F1-Score, accounting for a significant proportion across all categories, particularly in the semantic (32.26%) and discourse-level (26.39%) tasks. Accuracy is also widely reported, with notable usage in the lexical (18.52%) and discourse-level (25.00%) categories.

Precision and recall are consistently used across most levels, with their highest representation in the pragmatic category (24.00%). However, less frequently employed metrics, such as AUC and Jaccard Score, show minimal usage, highlighting a potential gap in evaluating ranking-based performance. The presence of unspecified metrics (e.g., 14.81% in lexical) suggests a degree of inconsistency in evaluation practices. The predominance of Precision and Recall metrics in the syntactic level, as shown in Figure 6, can be attributed to the nature of syntactic analysis tasks. Unlike semantic or discourse-level processing, syntactic-level figurative speech detection often relies on structural patterns and rule-based approaches, which benefit from evaluating how well the model retrieves relevant syntactic constructs (Recall) and how accurately it identifies them (Precision). Given that syntactic analysis typically involves well-defined linguistic features such as dependency relations, part-of-speech tags, and parse trees, precision and recall become crucial in assessing the correctness of pattern-matching algorithms and rule-based methods. Additionally, syntactic figurative language styles tend to exhibit clear structural cues, making these metrics highly relevant for evaluating model performance in capturing such patterns effectively.

The aggregated results in the **"All styles combined"** subplot reinforce the dominance of F1-Score and Accuracy, with other metrics such as Loss and Specificity receiving limited attention. These findings suggest that while classification-based metrics dominate the evaluation landscape, future research could benefit from a broader adoption of ranking and error-based measures to provide a more holistic assessment of model performance.

With this analysis in consideration, the remainder of our analysis will be focused on the 4 most used metrics: **Accuracy**, **F1-Score**, **Precision** and **Recall**

Table 7. Average performance metrics across figurative speech style levels.

Style Level	Accuracy	F1-Score	Precision	Recall
Lexical	78.67	75.70	71.00	79.10
Syntactic	91.00	65.67	58.60	52.80
Semantic	82.25	83.21	78.95	79.21
Discourse-level	88.36	88.06	88.14	88.69
Pragmatic	87.57	83.00	81.71	75.93

Performance Analysis Across Figurative Speech Style Levels Table 7 presents the performance of figurative speech detection models across different style levels using key evaluation metrics: accuracy, F1-score, precision, and recall. The results indicate that syntactic-level approaches achieve

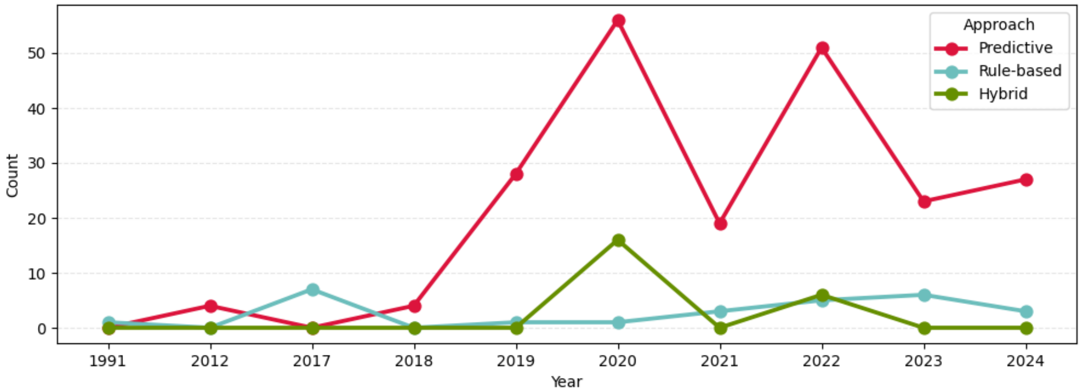

Figure 5. Temporal evolution of methodological approaches in figurative speech detection.

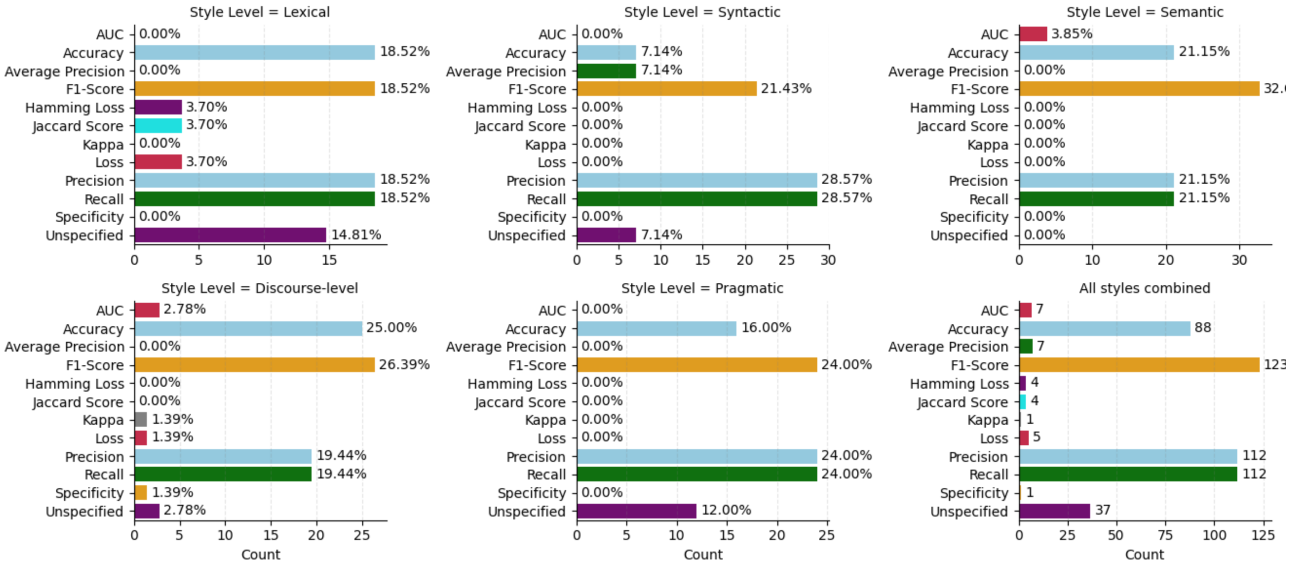

Figure 6. Distribution of evaluation metrics used in figurative speech detection studies across different linguistic style levels.

the highest accuracy (91.00%), but their relatively lower F1-score (65.67%) suggests potential class imbalance or difficulty in capturing structural patterns effectively.

Discourse-level figurative speech yields the best overall performance, with an F1-score of 88.06% and balanced precision (88.14%) and recall (88.69%), demonstrating the importance of contextual cues in improving model effectiveness.

Lexical-level approaches show moderate accuracy (78.67%) and F1-score (75.70%), indicating that surface-level features alone may not be sufficient to capture deeper figurative meanings. Pragmatic styles, while achieving a good balance of metrics, display lower recall (75.93%), implying challenges in generalization to diverse pragmatic contexts.

These findings highlight the varying complexities of figurative speech styles and suggest that the choice of feature extraction techniques plays a crucial role in achieving optimal performance. Table 8 presents the performance of various feature extraction techniques across different figurative speech processing levels.

The results demonstrate that embedding-based features consistently achieve high scores across all metrics and style levels, particularly in the lexical (Accuracy: 86.67%) and pragmatic (Accuracy: 92.83%) categories. Graph-based features exhibit strong performance in lexical and discourse-level styles, with high F1-scores of 96.00% and 86.67%, respectively. This indicates their effectiveness in capturing lexical relationships within figurative language expressions, leveraging external knowledge sources such as ConceptNet [49] and WordNet [50] to enhance the representation of semantic and relational dependencies.

Interestingly, occurrence-based features contribute significantly to semantic processing, achieving an accuracy of 79.33% and an F1-score of 83.50%. This suggests that tracking term frequency and distribution is crucial in understanding figurative meaning in context.

Despite the overall dominance of embedding features, the results highlight the complementary role of combining multiple feature types, as seen in the discourse-level and semantic categories, where hybrid feature sets yield the highest precision and recall values.

Future research should explore advanced feature selection techniques to optimize performance further and assess the trade-offs between feature complexity and computational efficiency.

Table 8. Performance comparison of feature extraction techniques across different figurative speech style levels. The table presents the accuracy, F1-score, precision, and recall for each feature type, highlighting their effectiveness in various figurative language processing tasks. The results show the dominance of embedding-based features across all style levels, while graph-based and occurrence-based features demonstrate their strengths in syntactic and semantic processing, respectively.

Metrics	Style level	Embedding	Figurative	Graph	Lexical	Misc	Morpho	Occurrence	Semantic	Sentiment	Syntactic
Accuracy	Lexical	86.67	-	96.00	-	-	-	-	-	-	58.00
	Syntactic	-	-	-	91.00	-	-	-	-	-	91.00
	Semantic	82.20	-	-	86.00	86.00	-	79.33	72.00	82.50	86.00
	Discourse-level	89.38	94.00	-	88.89	94.00	79.25	88.50	95.67	88.14	87.00
	Pragmatic	92.83	-	84.00	72.00	-	-	91.00	84.00	-	87.50
	All	85.83	85.83	85.83	85.83	85.83	85.83	85.83	85.83	85.83	85.83
F1-Score	Lexical	73.00	-	96.00	-	-	-	-	-	-	75.00
	Syntactic	97.00	-	-	-	-	-	-	-	-	-
	Semantic	80.22	86.67	86.67	85.75	85.75	86.67	83.50	84.00	84.43	85.83
	Discourse-level	87.44	88.50	86.67	90.00	85.14	86.62	89.00	91.17	87.45	87.94
	Pragmatic	82.00	-	79.00	-	-	-	89.00	79.00	-	84.00
	All	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62
Precision	Lexical	71.71	-	92.00	-	-	-	-	-	-	58.00
	Syntactic	97.00	-	41.00	-	-	-	-	-	-	-
	Semantic	77.77	-	-	87.00	87.00	-	81.50	76.00	76.00	-
	Discourse-level	90.56	95.00	-	88.50	84.50	86.40	87.67	94.50	87.25	86.67
	Pragmatic	86.25	-	85.00	67.00	-	67.00	92.00	85.00	-	81.33
	All	77.95	77.95	77.95	77.95	77.95	77.95	77.95	77.95	77.95	77.95
Recall	Lexical	73.71	-	95.00	-	-	-	-	-	-	90.00
	Syntactic	98.00	-	38.00	-	-	-	-	-	-	-
	Semantic	79.54	-	-	81.00	81.00	-	78.50	76.00	76.00	-
	Discourse-level	87.67	94.00	-	91.25	84.50	86.6	89.22	94.00	86.75	89.00
	Pragmatic	79.50	-	73.00	67.00	-	67.00	86.00	73.00	-	75.33
	All	77.56	77.56	77.56	77.56	77.56	77.56	77.56	77.56	77.56	77.56

Table 9. Correlation analysis between dataset size and performance metrics.

Metric	Pearson Correlation	Pearson p-value	Spearman Correlation	Spearman p-value
Accuracy	-0.1536	0.3376	-0.0054	0.9732
F1-Score	-0.0829	0.5148	0.1817	0.1507
Precision	0.0610	0.6837	0.1870	0.2081
Recall	-0.1159	0.4380	0.2050	0.1670

Impact of Dataset Size on Performance Table 9 presents the correlation analysis between dataset size and key performance metrics using Pearson and Spearman correlation coefficients. The results indicate weak correlations across all metrics, with the highest Spearman correlation observed for recall ($r_s = 0.2050$), though it remains statistically insignificant ($p = 0.1670$).

Interestingly, negative Pearson correlations for accuracy ($r_p = -0.1536$) and recall ($r_p = -0.1159$) suggest a potential decline in performance with increasing dataset size, possibly due to factors such as overfitting or increased noise in larger datasets. However, none of the computed correlations achieved statistical significance ($p > 0.05$), indicating that dataset size alone may not be a strong predictor of performance.

These findings suggest that while dataset size plays a role, additional factors such as data quality, feature richness, and model capacity should be considered to enhance figurative speech detection performance.

Which languages have seen particular advancements in this field ?

Analysis of Language Coverage in Figurative Speech Detection Figure 7 reveals a significant concentration of studies on the English language, with **42** references, far surpassing other languages. Arabic follows with **7** studies, while Spanish, Gujarati, French, German, and Bengali each have limited coverage with **2-3** references. Notably, several languages, including Chinese, Japanese, and Persian, are represented by only a single study. To streamline analysis, we categorize all languages with a single reference under the *low-resource* group, recognizing their limited exploration in the field.

It is important to note that these numbers reflect the search criteria and available datasets rather than an exhaustive representation of research efforts globally. Factors such as data availability, linguistic complexity, and regional research focus play a critical role in determining the prominence of specific languages in the literature. The predominance of English may be attributed to the abundance of publicly available datasets and its widespread use in academic and technological communities. Meanwhile, the relatively lower representation of other languages suggests potential gaps that warrant further investigation and resource development to support figurative speech detection in diverse linguistic settings.

Languages with only a single referenced work, such as Chinese, Hindi, and Japanese, can be categorized as *low-resource* within the context of figurative speech detection research. These languages face substantial challenges due to the lack of comprehensive datasets and linguistic tools, limiting advancements in this field. Notably, multilingual studies, represented by **2 references**, suggest an emerging interest in cross-linguistic comparisons and transfer learning approaches, which may help bridge the resource gap for underrepresented languages.

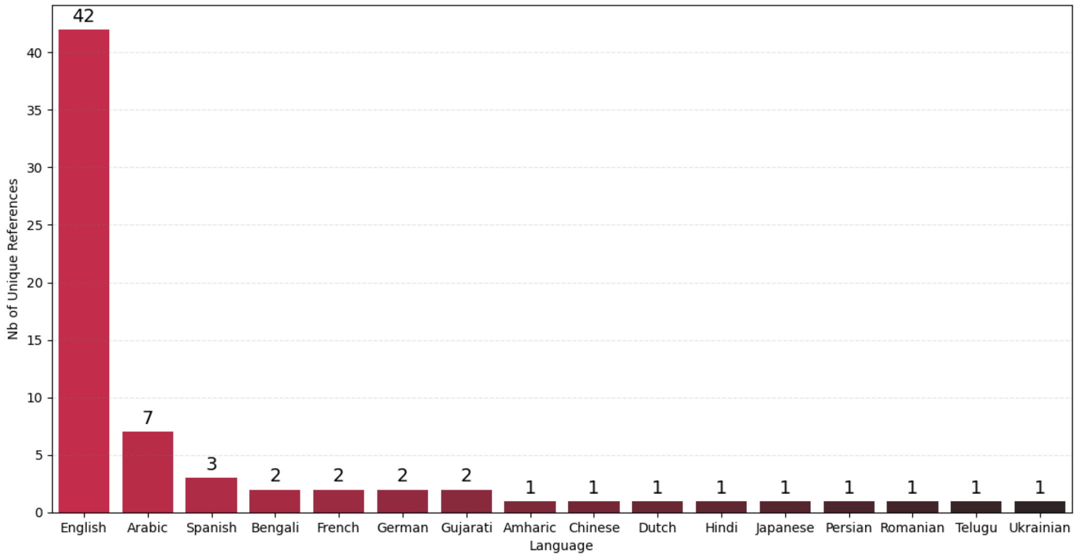

Figure 7. Distribution of figurative speech detection studies across languages.

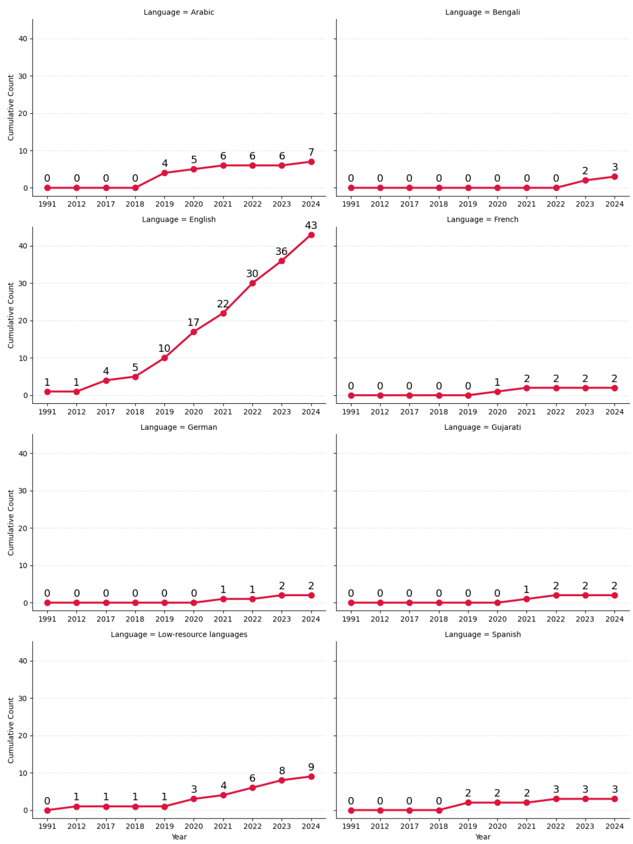

Figure 8. Research studies evolution of figurative speech detection studies across languages over time.

The temporal analysis in Figure 8 further reveals that English has experienced a sharp increase in research interest over the past decade, particularly post-2018, reflecting the rapid advancements in deep learning and the proliferation of benchmark datasets such as SemEval. Other languages, however, exhibit sporadic growth patterns, with Arabic and Spanish showing a gradual increase in recent years, whereas low-resource languages remain relatively stagnant, emphasizing the need for targeted initiatives to expand their representation.

These findings underscore the critical need for fostering research efforts beyond high-resource languages, encouraging the development of multilingual resources and methodologies to ensure more inclusive and equitable progress in figurative speech detection across diverse linguistic landscapes.

Dataset and External Resource Distribution Across Languages Table 10 presents a comprehensive overview of the distribution of available datasets and external resources across various languages and figurative speech style levels. The table highlights the disparity in resource availability, with English continuing to dominate, benefiting from a rich ecosystem of datasets and linguistic resources such as corpora and lexicons. Other languages, such as Arabic and Spanish, show a moderate level of resource availability but still lag behind in terms of diversity and coverage.

Low-resource languages exhibit a significant lack of both datasets and supporting resources, limiting their representation in figurative speech detection research. Syntactic-level studies stand out with a notable reliance on external resources such as linguistic databases and lexical knowledge bases *(e.g., BabelNet, Wiktionary)*, which compensate for the scarcity of annotated datasets. This observation underscores the critical role of linguistic resources in facilitating syntactic-level figurative speech analysis. Conversely, lexical and pragmatic style levels exhibit a more balanced use of both datasets and external resources.

The analysis suggests that future research efforts should focus on enhancing resource availability for underrepresented languages and style levels to achieve a

Table 10. Distribution of resource availability across languages and figurative style levels.

Style level	Arabic	Bengali	English	French	German	Low-resource languages	Spanish
Lexical	0	0	3	0	0	1	0
Syntactic	0	0	3	0	5	1	0
Semantic	2	0	7	1	0	2	4
Discourse-level	1	3	11	1	0	2	4
Pragmatic	3	0	9	0	0	1	0
Total	6	3	33	2	5	7	8

more comprehensive understanding of figurative speech detection across linguistic and cultural contexts.

Low-resource languages in figurative speech detection Table 10 and Figure 7 highlight the presence of low-resource languages within the domain of figurative speech detection. Languages such as Amharic, Chinese, Dutch, Hindi, Japanese, Persian, Romanian, Telugu, and Ukrainian are represented with only a single reference, indicating minimal research activity. Despite this, the inclusion of at least one dataset or resource for these languages signifies some level of attention. However, their limited representation underscores the significant gap in resources available for effective figurative speech detection.

The lack of datasets and supporting linguistic resources for these languages poses challenges in developing robust figurative speech models. Addressing these gaps would require targeted data collection efforts and cross-linguistic adaptation techniques to enhance coverage across diverse linguistic landscapes.

Analysis of Methodological Approaches Across Languages Figure 9 illustrates the distribution of methodological approaches used for figurative speech detection across various languages. The analysis reveals a significant dominance of predictive approaches, while rule-based and hybrid methods are relatively underrepresented. This trend highlights the growing reliance on machine learning techniques in the field, particularly in languages with abundant resources such as English.

Dominance of Predictive Approaches. Predictive methods, primarily driven by machine learning and deep learning models, constitute the majority of research efforts, with English leading at 32 references. This prevalence can be attributed to the widespread availability of annotated datasets, pre-trained models, and computational resources. Other languages, such as Arabic (6 references) and Spanish (3 references), also show a clear preference for predictive techniques, reflecting a broader trend in natural language processing research.

Rule-Based Methods and Hybrid Approaches. Rule-based methods are mainly observed in English (10 references), German (1 reference), and Gujarati (2 references), indicating their continued relevance in scenarios where handcrafted linguistic rules can complement data-driven methods. However, their overall limited presence suggests a shift away from traditional approaches in favor of automated learning systems. Hybrid methods, which integrate predictive and rule-based strategies, are scarcely employed, appearing only in low-resource languages with 2 references. This suggests that hybrid approaches may offer potential solutions for languages with limited labeled data.

The low-resource category, encompassing languages such as Amharic, Telugu, and Romanian, exhibits a relatively balanced distribution between predictive, rule-based, and hybrid approaches. This distribution underscores the challenges faced in low-resource settings, where researchers may resort to hybrid methodologies to leverage linguistic resources alongside data-driven techniques.

The disproportionate focus on predictive methods raises concerns about the generalizability of models across diverse linguistic contexts. Future research could explore the potential of hybrid approaches to bridge the gap in resource-scarce languages while also examining the transferability of predictive models to underrepresented linguistic domains.

Impact of Language Typology on Figurative Speech Detection Languages can generally be classified into two typological categories: *analytical* and *morphologically rich* languages (MRLs). Analytical languages, such as English, rely primarily on word order and auxiliary words to convey meaning, exhibiting minimal inflectional morphology. On the other hand, MRLs, such as Arabic, encode extensive grammatical information within words through affixes, conjugations, and other morphological processes. For instance, in English, pluralization is mostly achieved by adding an *"-s"* to a noun (e.g., *book* vs. *books*), whereas in Arabic, pluralization can involve internal vowel changes (e.g., كِتَابٌ vs. كُتُبٌ). Similarly, verb tenses in English often require auxiliary verbs (*I will go*), whereas Arabic uses rich verb conjugations that incorporate tense, subject, and aspect within a single word form (سَأَذْهَبُ). These structural differences necessitate distinct processing techniques, making rule-based approaches more viable for Arabic, while English benefits more from predictive models. This fundamental difference in structure has direct implications for figurative speech detection, as illustrated in Figure 10.

The analysis presented in Figure 10 reveals a clear preference for predictive methods across both language categories, with 76% of studies on analytical languages employing such approaches, compared to 68% for MRLs. This dominance can be attributed to the widespread availability of large corpora and pretrained language models, which perform well in analytical languages due to their simpler syntactic structure and greater availability of annotated resources. Additionally, the inherent complexity of MRLs poses challenges to purely predictive models, often

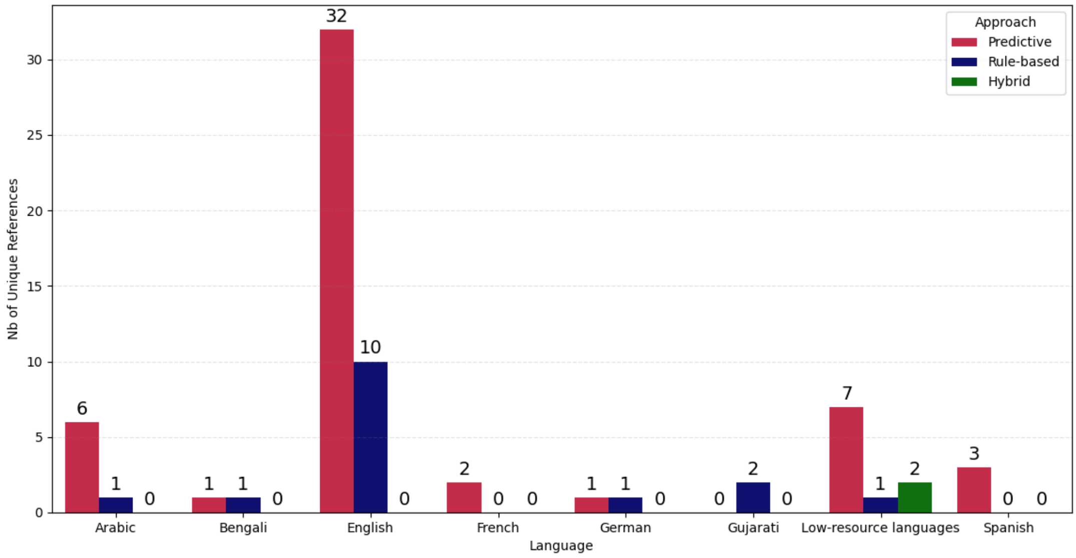

Figure 9. Figurative language identification approaches usage per language

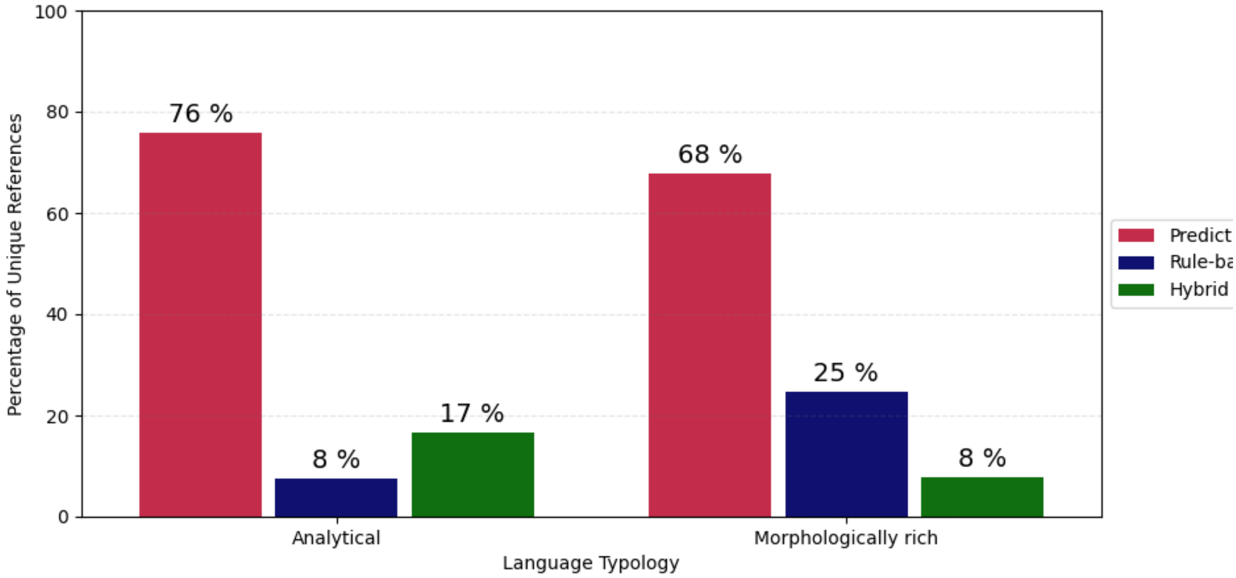

Figure 10. Comparison of figurative speech detection methods across language typologies.

requiring additional rule-based enhancements to capture intricate morphological patterns.

Despite the overall dominance of predictive methods, rule-based approaches are notably more prevalent in MRLs (25%) than in analytical languages (8%). This can be explained by the necessity of leveraging linguistic rules to address rich morphology and syntactic complexity that predictive models struggle to generalize effectively. Conversely, the lower proportion of hybrid methods in MRLs (8%) compared to analytical languages (17%) suggests that researchers may face challenges in integrating both approaches effectively, potentially due to resource limitations and the increased complexity of morphological analysis.

The distribution of methods across typologies highlights the need for a tailored approach to figurative speech detection, considering linguistic characteristics and data availability. While predictive models remain the preferred choice, a growing trend towards hybrid methods in analytical languages suggests that further research could explore similar integration strategies for MRLs to improve performance and generalization across diverse linguistic structures.

Discussion

Performance varies across different figurative speech style levels, with discourse-level tasks achieving the highest overall accuracy and F1-scores. This suggests that contextual cues available in discourse provide valuable information for figurative language detection. Syntactic-level tasks achieve high accuracy but lower F1-scores, indicating challenges in handling structural complexity and class imbalance. Pragmatic and semantic styles, which rely heavily on contextual understanding, exhibit variability in precision and recall, reflecting the challenges in capturing implicit meanings effectively.

The analysis also reveals a strong reliance on English datasets, highlighting a significant resource imbalance across languages. English datasets dominate the landscape due to their widespread availability, while low-resource languages such as Bengali and Gujarati remain underrepresented. The findings emphasize the need for more diverse and representative datasets to ensure equitable advancements in figurative speech detection across different linguistic contexts. Additionally, dataset availability varies by style level, with discourse-level figurative speech receiving the most attention due to the prevalence of social media data.

Embedding-based feature extraction techniques emerge as the most dominant approach, accounting for a substantial portion of the analyzed studies. This indicates a growing preference for leveraging pre-trained language models and distributed representations to capture contextual and semantic nuances of figurative language. However, the relatively low usage of figurative-specific features suggests that current approaches rely more on general linguistic features rather than specialized ones tailored for figurative speech. Syntactic and lexical features continue to play a crucial role, particularly for rule-based methods, which benefit from structural cues in figurative expressions.

Deep learning models dominate the methodological landscape, reflecting their superior capability to capture complex contextual relationships in figurative language. Traditional machine learning models, such as SVM and Random Forest, maintain relevance for structured and interpretable analyses but are significantly outpaced by deep learning approaches. Rule-based methods, though used less frequently, remain crucial for specific figurative styles with well-defined structures. The analysis also indicates a rising interest in ensemble models, combining various approaches to enhance robustness and accuracy.

Our findings largely align with existing literature in several key areas, while also revealing notable deviations that highlight evolving trends and emerging challenges in figurative speech detection.

Consistent with prior reviews, our analysis confirms the dominance of English datasets, reinforcing the notion that resource availability remains a major barrier for low-resource languages. Studies such as [51] and [52] similarly observed a disproportionate focus on English corpora, emphasizing the need for broader linguistic diversity. Moreover, the widespread adoption of deep learning techniques observed in our study mirrors previous findings that highlight the transition from traditional machine learning methods to more sophisticated neural architectures. The continued reliance on embeddings for feature extraction aligns with earlier observations, underscoring their effectiveness in capturing semantic richness across various figurative styles.

Our findings indicate a pronounced shift towards deep learning, with hybrid models gaining traction only in recent years. This shift suggests a growing preference for data-driven methods and the diminishing reliance on handcrafted features, which were once considered foundational in figurative speech detection. Additionally, our analysis reveals an increased focus on pragmatic and discourse-level figurative speech styles, a trend that was less evident in earlier studies which primarily concentrated on syntactic and lexical styles. This evolution can be attributed to the rise of social media platforms, where contextual and pragmatic understanding is crucial for effective interpretation.

An important trend observed in our findings, which diverges from prior studies, is the increasing integration of external resources such as lexicons, knowledge bases, and pretrained models. This suggests a growing recognition of the importance of domain adaptation and explainability, aspects that were previously overlooked in figurative speech detection. Furthermore, our findings highlight a shift towards more comprehensive evaluation metrics, moving beyond traditional accuracy-based assessments to include measures such as F1-score and precision-recall balance, offering a more nuanced understanding of model performance.

These comparisons underscore both the progress and persisting gaps in the field, reinforcing the need for further interdisciplinary efforts to bridge linguistic, computational, and cultural dimensions of figurative speech detection.

The findings of this study provide several key implications for future research and practical applications in figurative speech detection. These implications span across dataset development, methodological improvements, and real-world deployments.

Data Expansion and Diversity Our analysis underscores the pressing need for expanding datasets to include a more diverse range of languages, particularly low-resource languages. The limited representation of morphologically rich languages such as Arabic suggests that future research should prioritize the creation of annotated corpora that capture linguistic complexities beyond English. Additionally, efforts should focus on incorporating more varied sources, such as conversational data and literary texts, to enhance the generalizability of detection models across different contexts.

Methodological Advancements The observed dominance of deep learning approaches highlights the potential for leveraging advanced architectures such as transformers and hybrid models to improve figurative speech detection. While our analysis focused primarily on text-based figurative speech detection, it is important to acknowledge that there have been efforts in the multimodal domain, integrating textual, visual, and contextual information to achieve a richer representation of figurative meaning. Future research should further explore such multimodal approaches, which can offer deeper insights into figurative expressions across different modalities, enhancing model performance in complex real-world applications.

Practical Applications The evolving landscape of figurative speech detection presents promising applications across various domains. In social media analysis, improved detection systems can aid in identifying misinformation, sarcasm, and hate speech more accurately, thus contributing to safer online environments. In educational contexts, figurative speech detection tools can enhance language learning by offering automated feedback and insights into figurative expressions used in literature. Additionally, customer sentiment analysis in businesses can benefit from more nuanced detection capabilities, enabling brands to understand and respond to consumer feedback more effectively.

Overall, our study provides a roadmap for future research to enhance the robustness and the applicability of figurative speech detection models in both academic and industrial settings.

While this study provides a comprehensive analysis of figurative speech detection, several limitations must be acknowledged.

Data Availability and Diversity The analysis is limited by the availability and diversity of datasets, which may not fully represent the entire spectrum of figurative speech across different languages and cultural contexts. Low-resource languages, in particular, suffer from a lack of annotated data, which may introduce biases toward languages with richer textual resources, such as English.

Feature Engineering Biases Despite the effort to categorize feature engineering techniques, the effectiveness of these features may vary across datasets and figurative styles. The generalizability of findings across different domains is limited, as certain feature sets may be optimized for specific tasks but underperform in others.

Model-Centric Focus This study primarily focused on evaluating model performance, potentially overlooking practical deployment challenges such as computational efficiency, real-time processing constraints, and user interpretability. Additionally, the dominance of deep learning models in the literature may overshadow promising rule-based and hybrid approaches that could be more suitable for specific applications.

Subjectivity in Annotation The inherent subjectivity in annotating figurative speech poses a challenge in achieving consistent and reliable datasets. Variability in annotation guidelines and inter-annotator agreement could impact the reported model performance and introduce inconsistencies across different studies.

Exclusion of Multimodal Approaches While our analysis focused on text-based detection, emerging multimodal approaches that integrate visual and auditory cues were not considered. This exclusion may limit the applicability of our findings to domains where figurative expressions rely heavily on non-textual elements.

Despite these limitations, the insights gained from this study provide a valuable foundation for future research and practical applications in figurative speech detection.

Building upon the insights gained from this study, several promising directions for future research in figurative speech detection can be explored.

Expansion of Multilingual and Low-Resource Studies Future research should focus on developing datasets and models tailored for underrepresented languages, particularly morphologically rich and low-resource languages. Leveraging transfer learning, cross-lingual embeddings, and multilingual pre-trained models such as mBERT and XLM-RoBERTa can help bridge the linguistic gap and enhance performance across diverse languages.

Integration of Multimodal Approaches Given the inherent complexity of figurative speech, integrating multimodal signals such as images, videos, and audio alongside text can provide deeper contextual understanding. Future studies should explore how multimodal fusion techniques can improve the detection of figurative expressions, particularly in domains such as social media and marketing.

Improved Explainability and Interpretability As deep learning models become more prevalent in figurative speech detection, enhancing their explainability remains crucial. Future research should focus on incorporating explainable AI (XAI) techniques such as attention visualization, counterfactual reasoning, and post-hoc explanations to provide clearer insights into model decisions and increase trustworthiness.

Task-Specific Model Optimization Different figurative speech styles (e.g., metaphor, parallelism, litotes) pose unique challenges that may require specialized models. Future work could focus on developing tailored architectures and training strategies optimized for specific styles, allowing for more accurate and efficient detection.

Robustness and Generalization Across Domains Ensuring that models generalize well to unseen domains and contexts is critical for practical applications. Future studies should investigate domain adaptation techniques and adversarial training to improve the robustness of figurative speech detection models across varied data sources and genres.

Ethical Considerations and Bias Mitigation Future research should address ethical challenges related to fairness and bias in figurative speech detection. Efforts to ensure balanced datasets, mitigate cultural biases, and enhance model transparency will contribute to the responsible deployment of these technologies in real-world scenarios.

By pursuing these directions, researchers can further advance the field of figurative speech detection, making it more inclusive, interpretable, and applicable across various domains and languages.

Conclusion

This systematic literature review provides a comprehensive examination of figurative speech detection, analyzing the datasets, methods, and challenges that characterize the field. Our findings underscore the dominance of English datasets, highlighting the urgent need for greater linguistic diversity and representation of low-resource languages. The imbalance in dataset availability suggests that future efforts should prioritize the development of annotated corpora for morphologically rich and underrepresented languages to ensure the broader applicability of detection models.

From a methodological perspective, our analysis reveals a clear shift toward deep learning approaches, with transformer-based models such as BERT and RoBERTa leading the field. While these models achieve state-of-the-art performance, traditional machine learning approaches and hybrid methods continue to play a role in resource-constrained and interpretability-driven scenarios. Feature engineering techniques, particularly embedding-based representations, have proven to be instrumental in improving model performance, though there remains room for incorporating more figurative-specific features to capture the nuances of different styles of figurative speech.

The review also identifies significant variations in performance across different figurative speech styles and linguistic levels. Discourse-level figurative speech, largely driven by the availability of social media data, has received the most attention, whereas syntactic and pragmatic styles present unique challenges that require further exploration. The variability in evaluation metrics across studies highlights the need for standardization in performance assessment to facilitate fair comparisons and reproducibility.

Despite advancements, key challenges persist in figurative speech detection. These include annotation inconsistencies, class imbalances, and model interpretability issues. The reliance on English-centric benchmarks and the limited focus on low-resource languages further exacerbate these challenges. Additionally, our study highlights the exclusion of multimodal approaches, which, while beyond the scope of our analysis, present a promising avenue for future research by integrating textual, visual, and contextual information.

Looking forward, we propose several directions to advance the field. Expanding multilingual research, developing explainable AI models, and improving the generalizability of models across domains are crucial steps to enhancing the robustness and inclusivity of figurative speech detection systems. Furthermore, ethical considerations, including bias mitigation and fairness in datasets and models, should remain a priority to ensure responsible AI deployment.

In conclusion, while significant progress has been made in figurative speech detection, there remain numerous opportunities for growth and refinement. This study provides a foundation for future research, aiming to bridge existing gaps and support the development of more effective, inclusive, and interpretable solutions for understanding figurative language across diverse linguistic and cultural contexts.

Acknowledgements

This class file was developed by Sunrise Setting Ltd, Brixham, Devon, UK.
Website: http://www.sunrise-setting.co.uk

References

[1] H. Tanriverdiyeva. "On the role of metaphor in enriching the lexical layer of language". In: *Scientific Works* (Apr. 22, 2024). DOI: 10.62706/bqiz.2024.v18i1.133.

[2] Yi-Chen Chen. "Assessing the lexical richness of figurative expressions in Taiwanese EFL learners' writing". In: *Assessing Writing* 43 (2020), p. 100414. DOI: 10.1016/J.ASW.2019.06.007.

[3] "Lexical Stylistic Devices and Literary Terms of Figurative Language". In: *International Journal of Recent Technology and Engineering* (Oct. 22, 2019). DOI: 10.35940/ijrte.c1054.1083s19.

[4] Aasma Younas et al. "An Analysis of Figurative Language in Taufiq Rafat's Poem "The Time to Love"". In: *Qlantic Journal of Social Sciences and Humanities* (Mar. 30, 2024). DOI: 10.55737/qjssh.379019364.

[5] Li Shaofan. "Distinctions on several confused figurative speeches". In: (2003).

[6] T. Volkova and O. V. Orlova. "Language personality of suicident: linguodiscursive markers of ambivalence and demonstrability". In: *Sibirskiy filologicheskiy zhurnal* (2022). DOI: 10.17223/18137083/78/22.

[7] M. Dyer. "Comprehension and Acquisition of Figurative Expressions With Phrasal/ Lexical Memory". In: *Metaphor and Symbol* 4 (Sept. 1, 1989), pp. 173–201. DOI: 10.1207/S15327868MS0403_5.

[8] S. Chandler. "Metaphor Comprehension: A Connectionist Approach to Implications for the Mental Lexicon". In: *Metaphor and Symbol* 6 (Dec. 1, 1991), pp. 227–258. DOI: 10.1207/S15327868MS0604_1.

[9] L. A. Mustafayeva. "Some Facts of the Usages of Rhetorical Devices (Zeugma) In Speech and Writing". In: *International Journal of Social Science And Human Research* (2023). DOI: 10.47191/ijsshr/v6-i1-90.

[10] John B. Anderson. "The Grammar of Figurativeness, with Illustrations Drawn from English". In: *English Studies* 95 (2014), pp. 971–989. DOI: 10.1080/0013838X.2014.962304.

[11] R. Carston and C. Wearing. "Hyperbolic language and its relation to metaphor and irony". In: *Journal of Pragmatics* 79 (2015), pp. 79–92. DOI: 10.1016/J.PRAGMA.2015.01.011.

[12] C. Burgers, K. Y. R. D. Lavalette, and G. Steen. "Metaphor, hyperbole, and irony: Uses in isolation and in combination in written discourse". In: *Journal of Pragmatics* 127 (2018), pp. 71–83. DOI: 10.1016/J.PRAGMA.2018.01.009.

[13] Sri Arfani and Muhammad Rino Adi Laksono. "Figurative Language in "The Beginning After The End" Volume 10 Novel by Turtleme". In: *JELL (Journal of English Language and Literature) STIBA-IEC Jakarta* (2024). DOI: 10.37110/jell.v9i02.241.

[14] R. Tambunan. "AN ANALYSIS OF FIGURATIVE MEANING ON OBAMA'S PRESIDENTIAL SPEECH". In: (2015).

[15] Pegy Eka Pertiwi and Nurma Dhona Handayani. "Semantic Analysis of Figurative Language Found in Nikkie Tutorials Product Review". In: *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature* (2023). DOI: 10.24256/ideas.v11i2.3505.

[16] Afrianto Afrianto. "A Semantic Analysis of Figurative Language in Michael Jackson's Selected Song Lyrics". In: 13 (2016).

[17] Wu Yuan. "Figures of Speech in College English (Intensive Reading)". In: *Journal of Shanghai Institute & Electric Power* (2002).

[18] Alessandro Aru. "Conceptual Blending and Compound Figures of Speech". In: *International Journal of English Linguistics* (2024). DOI: 10.5539/ijel.v14n4p45.

[19] Richard M. Roberts and Roger Kreuz. "Why Do People Use Figurative Language?" In: *Psychological Science* 5 (1994), pp. 159–163. DOI: 10.1111/j.1467-9280.1994.tb00653.x.

[20] C. Burgers, E. Konijn, and G. Steen. "Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony". In: *Communication Theory* 26 (2016), pp. 410–430. DOI: 10.1111/COMT.12096.

[21] Alina Ruslanivna Shvachko. "Functions of irreverent images in English Poetry". In: (2024).

[22] R. Harris, Brian M. Friel, and Nolan Rett Mickelson. "Attribution of discourse goals for using concrete- and abstract-tenor metaphors and similes with or without discourse context". In: *Journal of Pragmatics* 38 (2006), pp. 863–879. DOI: 10.1016/J.PRAGMA.2005.06.010.

[23] A. Musolff. "Metaphor, irony and sarcasm in public discourse". In: *Journal of Pragmatics* 109 (2017), pp. 95–104. DOI: 10.1016/J.PRAGMA.2016.12.010.

[24] Jie Tan. "Figurative Units in the Communicative Practices of Mediatized Political Discourse". In: *Scientific Research and Development. Modern Communication Studies* (2024). DOI: 10.12737/2587-9103-2024-13-2-52-59.

[25] M. L. Mele. "Kanuri proverbs: metaphoric conceptualization of a cultural discourse". In: *Journal of African Cultural Studies* 25 (2013), pp. 333–348. DOI: 10.1080/13696815.2012.749783.

[26] Jonathan Marcos Macaraeg, J. R. O. Pelila, and Joselito C. Gutierrez. "Mirroring academic journeys through figurative language in the Book of Proverbs". In: *Cambodian Journal of Educational Research* (2024). DOI: 10.62037/cjer.2024.04.01.05.

[27] Li Qing. "A Pragmatic Study on English Euphemism and Its Implication to English Teaching". In: *Journal of Weifang University* (2008).

[28] LU Li-chun. "An Analysis of English Euphemism in Pragmatics Perspective". In: *Yinshan Academic Journal* (2007).

[29] A. Jaubert. "Dire et plus ou moins dire. Analyse pragmatique de l'euphémisme et de la litote". In: *Langue Francaise* 160 (2008), pp. 105–116. DOI: 10.3917/LF.160.0105.

[30] Ben Wetherbee. "Places in the polity of rhetoric : topoi, evolution, and the fragmentation of discourse." In: (2015). DOI: 10.18297/ETD/2325.

[31] Wei Hong-xia. "On the Rhetoric under the Perspective of Pragmatics". In: *Journal of Xichang College* (2011).

[32] Apurbalal Senapati. "Oxymoron: An Automatic Detection from the Corpus". In: *Procedia Computer Science*. 6th International Conference on AI in Computational Linguistics 244 (Jan. 2024), pp. 49–56. ISSN: 1877-0509. DOI: 10.1016/j.procs.2024.10.177. (Visited on 01/09/2025).

[33] Jatin C. Modh and Jatinderkumar R. Saini. "Dynamic Phrase Generation for Detection of Idioms of Gujarati Language Using Diacritics and Suffix-based Rules". In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 12.7 (Sept. 2021). ISSN: 2156-5570. DOI: 10.14569/IJACSA.2021.0120728. (Visited on 01/07/2025).

[34] Olena Levchenko, O. Tyshchenko, and M. Dilai. "Automated Identification of Metaphors in Annotated Corpus(Based on Substance Terms)". In: *International Conference on Computational Linguistics and Intelligent Systems*. 2021. (Visited on 01/25/2025).

[35] Shubhangi Rastogi and Divya Bansal. "Disinformation Detection on Social Media: An Integrated Approach". In: *Multimedia Tools and Applications* 81.28 (Nov. 2022), pp. 40675–40707. ISSN: 1573-7721. DOI: 10.1007/s11042-022-13129-y. (Visited on 01/07/2025).

[36] Hamada Nayel et al. "BENHA@IDAT: Improving Irony Detection in Arabic Tweets Using Ensemble Approach". In: *Fire*. 2019. (Visited on 01/25/2025).

[37] Anduamlak Abebe Fenta and Seffi Gebeyehu. "Automatic Idiom Identification Model for Amharic Language". In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22.8 (Aug. 2023), 210:1–210:9. ISSN: 2375-4699. DOI: 10.1145/3606864. (Visited on 01/07/2025).

[38] Takehiro Teraoka et al. "Automatic Detection of Metonymies Using Associative Relations between Words". In: *Cognitive Science* (2012). (Visited on 01/25/2025).

[39] Radu Tudor Ionescu and Adrian Gabriel Chifu. "FreSaDa: A French Satire Data Set for Cross-Domain Satire Detection". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. July 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533661. (Visited on 01/09/2025).

[40] Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. "Automatic Sarcasm Detection: Systematic Literature Review". In: *Information* 13.8 (Aug. 2022), p. 399. ISSN: 2078-2489. DOI: 10.3390/info13080399. (Visited on 01/09/2025).

[41] Jerry Liu et al. "Metaphor Detection Using Contextual Word Embeddings From Transformers". In: *Proceedings of the Second Workshop on Figurative Language Processing*. Ed. by Beata Beigman Klebanov et al. Online: Association for Computational Linguistics, July 2020, pp. 250–255. DOI: 10.18653/v1/2020.figlang-1.34. (Visited on 01/09/2025).

[42] Qusai Abuein et al. "ArSa-Tweets: A Novel Arabic Sarcasm Detection System Based on Deep Learning Model". In: *Heliyon* 10.17 (Sept. 2024), e36892. ISSN: 24058440. DOI: 10.1016/j.heliyon.2024.e36892. (Visited on 01/07/2025).

[43] U. Naseem et al. "Towards Improved Deep Contextual Embedding for the Identification of Irony and Sarcasm". In: *Proceedings of the International Joint Conference on Neural Networks*. 2020. DOI: 10.1109/IJCNN48605.2020.9207237.

[44] Reynier Ortega-Bueno, Paolo Rosso, and José E. Medina Pagola. "Multi-View Informed Attention-Based Model for Irony and Satire Detection in Spanish Variants". In: *Knowledge-Based Systems* 235 (Jan. 2022), p. 107597. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.107597. (Visited on 01/09/2025).

[45] Dawei Zhu et al. "ConFiguRe: Exploring Discourse-level Chinese Figures of Speech". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3374–3385. (Visited on 01/25/2025).

[46] Santosh Kumar Bharti, Reddy Naidu, and Korra Sathya Babu. "Hyperbolic Feature-based Sarcasm Detection in Telugu Conversation Sentences". In: *Journal of Intelligent Systems* 30.1 (Jan. 2021), pp. 73–89. ISSN: 2191-026X. DOI: 10.1515/jisys-2018-0475. (Visited on 01/07/2025).

[47] Julia Birke and Anoop Sarkar. "A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language". In: *Conference of the European Chapter of the Association for Computational Linguistics*. 2006.

[48] Cynthia Van Hee, Els Lefever, and Véronique Hoste. "SemEval-2018 Task 3: Irony Detection in English Tweets". In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. Ed. by Marianna Apidianaki et al. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 39–50. DOI: 10.18653/v1/S18-1005. URL: https://aclanthology.org/S18-1005/.

[49] Robyn Speer, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[50] George A Miller et al. "Introduction to WordNet: An on-line lexical database". In: *International journal of lexicography* 3.4 (1990), pp. 235–244.

[51] Ramona Kühn, Jelena Mitrović, and Michael Granitzer. *Computational Approaches to the Detection of Lesser-Known Rhetorical Figures: A Systematic Survey and Research Challenges*. June 2024. DOI: 10.48550/arXiv.2406.16674. arXiv: 2406.16674 [cs]. (Visited on 01/09/2025).

[52] Muhammad Abulaish, Ashraf Kamal, and Mohammed J. Zaki. "A Survey of Figurative Language and Its Computational Detection in Online Social Networks". In: *ACM Transactions on the Web* 14.1 (Feb. 2020), pp. 1–52. ISSN: 1559-1131, 1559-114X. DOI: 10.1145/3375547. (Visited on 01/09/2025).

[53] Rasikh Ali et al. "Deep Learning for Sarcasm Identification in News Headlines". In: *Applied Sciences* 13.9 (Jan. 2023), p. 5586. ISSN: 2076-3417. DOI: 10.3390/app13095586. (Visited on 01/07/2025).

[54] Dalia H. Elkamchouchi et al. "Hosted Cuckoo Optimization Algorithm with Stacked Autoencoder-Enabled Sarcasm Detection in Online Social Networks". In: *Applied Sciences* 12.14 (Jan. 2022), p. 7119. ISSN: 2076-3417. DOI: 10.3390/app12147119. (Visited on 01/07/2025).

[55] Kevin Alex Mathews and M. Strube. "Impact of Target Word and Context on End-to-End Metonymy Detection". In: *ArXiv* (Dec. 2021). (Visited on 01/25/2025).

[56] John Lawrence, Jacky Visser, and Chris Reed. "Harnessing Rhetorical Figures for Argument Mining: A Pilot Study in Relating Figures of Speech to Argument Structure". In: *Argument & Computation* 8.3 (Jan. 2017). Ed. by Randy Allen Harris and Chrysanne Di Marco, pp. 289–310. ISSN: 1946-2166, 1946-2174. DOI: 10.3233/aac-170026. (Visited on 01/14/2025).

[57] Dan Fass. "Met*: A Method for Discriminating Metonymy and Metaphor by Computer". In: *Computational Linguistics* 17.1 (1991), pp. 49–90. (Visited on 01/14/2025).

[58] D. Olaniyan et al. "Utilizing an Attention-Based LSTM Model for Detecting Sarcasm and Irony in Social Media". In: *Computers* 12.11 (2023). DOI: 10.3390/computers12110231.

[59] K.V.T.K.N. Prashanth and T. Ramakrishnudu. "Sarcasm-Based Tweet-Level Stress Detection". In: *Expert Systems* 41.4 (2024). DOI: 10.1111/exsy.13534.

[60] Michel Schwab, Robert Jäschke, and Frank Fischer. ""The Rodney Dangerfield of Stylistic Devices": End-to-End Detection and Extraction of Vossian Antonomasia Using Neural Networks". In: *Frontiers in Artificial Intelligence* 5 (June 2022). ISSN: 2624-8212. DOI: 10.3389/frai.2022.868249. (Visited on 01/07/2025).

[61] Sanzana Karim Lora et al. "A Transformer-Based Generative Adversarial Learning to Detect Sarcasm from Bengali Text with Correct Classification of Confusing Text". In: *Heliyon* 9.12 (Dec. 2023), e22531. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e22531. (Visited on 01/07/2025).

[62] Md Saifullah Razali et al. "Sarcasm Detection Using Deep Learning With Contextual Features". In: *IEEE Access* 9 (2021), pp. 68609–68618. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3076789. (Visited on 01/09/2025).

[63] Le Hoang Son et al. "Sarcasm Detection Using Soft Attention-Based Bidirectional Long Short-Term Memory Model With Convolution Network". In: *IEEE Access* 7 (2019), pp. 23319–23328. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2899260. (Visited on 01/07/2025).

[64] D. Šandor and M. Bagić Babac. "Sarcasm Detection in Online Comments Using Machine Learning". In: *Information Discovery and Delivery* 52.2 (2024), pp. 213–226. DOI: 10.1108/IDD-01-2023-0002.

[65] José Ángel González, Lluís-F. Hurtado, and Ferran Pla. "Transformer Based Contextualization of Pre-Trained Word Embeddings for Irony Detection in Twitter". In: *Information Processing & Management* 57.4 (July 2020), p. 102262. ISSN: 03064573. DOI: 10.1016/j.ipm.2020.102262. (Visited on 01/13/2025).

[66] Shiwei Zhang et al. "Irony Detection via Sentiment-Based Transfer Learning". In: *Information Processing & Management* 56.5 (Sept. 2019), pp. 1633–1644. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2019.04.006. (Visited on 01/09/2025).

[67] Jatin C. Modh, Jatinderkumar R. Saini, and Ketan Kotecha. "A Novel Morphological Analysis Based Approach for Dynamic Detection of Inflected Gujarati Idioms". In: *International Journal of Advanced Computer Science and Applications* 13.4 (2022). ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2022.0130422. (Visited on 01/07/2025).

[68] Arifur Rahaman et al. "Sarcasm Detection in Tweets: A Feature-based Approach Using Supervised Machine Learning Models". In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 12.6 (2021). ISSN: 2156-5570. DOI: 10.14569/IJACSA.2021.0120651. (Visited on 01/07/2025).

[69] Sameer Naser Olimat. "Euphemism in the Qur'an: A Corpus-based Linguistic Approach". In: *International Journal of Computational Linguistics (IJCL)* 10.2 (2019), pp. 16–32.

[70] B. S. Harish and Keerthi Kumar. "Automatic Irony Detection Using Feature Fusion and Ensemble Classifier". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 5.Regular Issue (2019), pp. 70–79. ISSN: 1989-1660. (Visited on 01/07/2025).

[71] Ashwini Joshi et al. "Sarcasm Detection Using Naïve Bayes SVM Hybrid". In: *International Journal of Recent Technology and Engineering (IJRTE)* 8.3 (Sept. 2019), pp. 1138–1142. ISSN: 22773878. DOI: 10.35940/ijrte.C4258.098319. (Visited on 01/07/2025).

[72] Rupali Amit Bagate and Ramadass Suguna. "Sarcasm Detection on Text for Political Domain— An Explainable Approach". In: *International Journal on Recent and Innovation Trends in Computing and Communication* 10.2s (Dec. 2022), pp. 255–268. ISSN: 2321-8169. DOI: 10.17762/ijritcc.v10i2s.5942. (Visited on 01/07/2025).

[73] Partha Mukherjee et al. "NegAIT: A New Parser for Medical Text Simplification Using Morphological, Sentential and Double Negation". In: *Journal of Biomedical Informatics* 69 (May 2017), pp. 55–62. ISSN: 15320464. DOI: 10.1016/j.jbi.2017.03.014. (Visited on 01/14/2025).

[74] Mohammed M. AL-Ani, Nazlia Omar, and Ahmed Adil Nafea. "A Hybrid Method of Long Short-Term Memory and Auto-Encoder Architectures for Sarcasm Detection". In: *Journal of Computer Science* 17.11 (Dec. 2021), pp. 1093–1098. ISSN: 1552-6607. DOI: 10.3844/jcssp.2021.1093.1098. (Visited on 01/07/2025).

[75] S. Farhan, R. Shoukat, and A. Aslam. "Automatic Sarcasm Detection on Cross-Platform Social Media Datasets: A GLoVe and Bi-LSTM Based Approach". In: *Journal of Universal Computer Science* 30.5 (2024), pp. 674–693. DOI: 10.3897/jucs.104790.

[76] Aaron Maladry et al. "The Limitations of Irony Detection in Dutch Social Media". In: *Language Resources and Evaluation* 58.4 (Dec. 2024), pp. 1355–1386. ISSN: 1574-0218. DOI: 10.1007/s10579-023-09656-1. (Visited on 01/09/2025).

[77] Vithyatheri Govindan and Vimala Balakrishnan. "INVESTIGATING THE IMPORTANCE OF HYPERBOLES TO DETECT SARCASM USING MACHINE LEARNING TECHNIQUES". In: *Malaysian Journal of Computer Science* 37.1 (Jan. 2024), pp. 71–88. ISSN: 0127-9084. DOI: 10.22452/mjcs.vol37no1.3. (Visited on 01/06/2025).

[78] S. Abarna, J.I. Sheeba, and S. Pradeep Devaneyan. "An Ensemble Model for Idioms and Literal Text Classification Using Knowledge-Enabled BERT in Deep Learning". In: *Measurement: Sensors* 24 (Dec. 2022), p. 100434. ISSN: 26659174. DOI: 10.1016/j.measen.2022.100434. (Visited on 01/11/2025).

[79] Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. "A Transformer-based Approach to Irony and Sarcasm Detection". In: *Neural Computing and Applications* 32.23 (Dec. 2020), pp. 17309–17320. ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-020-05102-3. arXiv: 1911.10401 [cs]. (Visited on 01/07/2025).

[80] Deepak Kumar Jain, Akshi Kumar, and Saurabh Raj Sangwan. "TANA: The Amalgam Neural Architecture for Sarcasm Detection in Indian Indigenous Language Combining LSTM and SVM with Word-Emoji Embeddings". In: *Pattern Recognition Letters* 160 (Aug. 2022), pp. 11–18. ISSN: 01678655. DOI: 10.1016/j.patrec.2022.05.026. (Visited on 01/07/2025).

[81] Jitesh Pradhan et al. "An Efficient Sarcasm Detection Using Linguistic Features and Ensemble Machine Learning". In: *Procedia Computer Science* 235 (2024), pp. 1058–1067. ISSN: 18770509. DOI: 10.1016/j.procs.2024.04.100. (Visited on 01/09/2025).

[82] Sudarshan S. Sonawane and Satish R. Kolhe. "TCSD: Term Co-occurrence Based Sarcasm Detection from Twitter Trends". In: *Procedia Computer Science* 167 (2020), pp. 830–839. ISSN: 18770509. DOI: 10.1016/j.procs.2020.03.422. (Visited on 01/09/2025).

[83] Kelsey Hilton, Akbar Siami Namin, and Keith S. Jones. "Metaphor Identification in Cybersecurity Texts: A Lightweight Linguistic Approach". In: *SN Applied Sciences* 4.2 (Jan. 2022), p. 60. ISSN: 2523-3971. DOI: 10.1007/s42452-022-04939-8. (Visited on 01/07/2025).

[84] Arvi Hurskainen. "Three Methods for Retrieving Proverbs: Application to English Language". In: *Technical Reports on Language Technology* (2024). DOI: 10.13140/RG.2.2.27209.15208. (Visited on 01/07/2025).

[85] Elizaveta G. Grishechko. "Language and Cognition behind Simile Construction: A Python-powered Corpus Research". In: *Training, Language and Culture* 7.2 (June 2023), pp. 80–92. ISSN: 25202073, 2521442X. DOI: 10.22363/2521-442X-2023-7-2-80-92. (Visited on 01/07/2025).

[86] Won Ik Cho et al. “Detecting Oxymoron in a Single Statement”. In: *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. Nov. 2017, pp. 1–5. DOI: 10.1109/ICSDA.2017.8384447. (Visited on 01/25/2025).

[87] Bilal Ghanem et al. “Irony Detection in a Multilingual Context”. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Cham: Springer International Publishing, 2020, pp. 141–149. ISBN: 978-3-030-45442-5. DOI: 10.1007/978-3-030-45442-5_18.

[88] N. Green and Larry Joshua Crotts. “Towards Automatic Detection of Antithesis”. In: *CMNA@COMMA*. 2020. (Visited on 01/25/2025).

[89] Reynier Ortega Bueno et al. “Overview of the Task on Irony Detection in Spanish Variants”. In: *IberLEF@SEPLN*. 2019. (Visited on 01/25/2025).

[90] M. Karp, N. Kunanets, and Yuliia Kucher. “Meiosis and Litotes in The Catcher in the Rye by Jerome David Salinger: Text Mining”. In: *International Conference on Computational Linguistics and Intelligent Systems*. 2021. (Visited on 01/25/2025).

[91] Sebastian-Vasile Echim et al. “Adversarial Capsule Networks for Romanian Satire Detection and Sentiment Analysis”. In: *Natural Language Processing and Information Systems*. Ed. by Elisabeth Métais et al. Cham: Springer Nature Switzerland, 2023, pp. 428–442. ISBN: 978-3-031-35320-8. DOI: 10.1007/978-3-031-35320-8_31.

[92] B. Talafha et al. “Sarcasm Detection and Quantification in Arabic Tweets”. In: *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*. Vol. 2021-November. 2021, pp. 1121–1125. DOI: 10.1109/ICTAI52525.2021.00177.

[93] Ramona Kuehn, Jelena Mitrović, and Michael Granitzer. “Hidden in Plain Sight: Can German Wiktionary and Wordnets Facilitate the Detection of Antithesis?” In: *Proceedings of the 12th Global Wordnet Conference*. Ed. by German Rigau, Francis Bond, and Alexandre Rademaker. University of the Basque Country, Donostia - San Sebastian, Basque Country: Global Wordnet Association, Jan. 2023, pp. 106–116. (Visited on 01/09/2025).

[94] Enrica Troiano et al. “A Computational Exploration of Exaggeration”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3296–3304. DOI: 10.18653/v1/D18-1367. (Visited on 01/14/2025).

[95] Haonan Li et al. “Target Word Masking for Location Metonymy Resolution”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3696–3707. DOI: 10.18653/v1/2020.coling-main.330. (Visited on 01/14/2025).

[96] Peratham Wiriyathammabhum. “TEDB System Description to a Shared Task on Euphemism Detection 2022”. In: *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. Ed. by Debanjan Ghosh et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 1–7. DOI: 10.18653/v1/2022.flp-1.1. (Visited on 01/25/2025).

[97] Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. “A Prompt Based Approach for Euphemism Detection”. In: *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. Ed. by Debanjan Ghosh et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 8–12. DOI: 10.18653/v1/2022.flp-1.2. (Visited on 01/09/2025).

[98] Felix Schneider et al. “Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features”. In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by Stefania Degaetano-Ortlieb et al. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021, pp. 96–100. DOI: 10.18653/v1/2021.latechclfl-1.11. (Visited on 01/14/2025).

[99] Nina Schneidermann, Daniel Hershcovich, and Bolette Pedersen. “Probing for Hyperbole in Pre-Trained Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Ed. by Vishakh Padmakumar, Gisela Vallejo, and Yao Fu. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 200–211. DOI: 10.18653/v1/2023.acl-srw.30. (Visited on 01/09/2025).

[100] Yuxue Hu et al. “Euphemism Identification via Feature Fusion and Individualization”. In: *Proceedings of the ACM Web Conference 2024*. WWW ’24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 2383–2394. ISBN: 9798400701719. DOI: 10.1145/3589334.3645433. (Visited on 01/25/2025).

[101] Preni Golazizian et al. “Irony Detection in Persian Language: A Transfer Learning Approach Using Emoji Prediction”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 2839–2845. ISBN: 979-10-95546-34-4. (Visited on 01/09/2025).

[102] Chiyu Zhang and Muhammad Abdul-Mageed. “Multi-Task Bidirectional Transformer Representations for Irony Detection”. In: *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*. Ed. by Parth Mehta et al. Vol. 2517. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 391–400. (Visited on 01/25/2025).

[103] Tharindu Ranasinghe et al. “RGCL at IDAT: Deep Learning Models for Irony Detection in Arabic Language”. In: *Fire*. 2019. (Visited on 01/25/2025).

[104] Kevalkumar Paida. “Double Negation Detection”. PhD thesis. Faculty of Computer Science and Mathematics: University of Passau, 2019.

[105] A. Alqahtani, A. Alsheddi, and L. Alhenaki. “Text-Based Sarcasm Detection on Social Networks: A Systematic Review”. In: *International Journal of Advanced Computer Science and Applications* 14.3 (2023), pp. 313–328. DOI: 10.14569/IJACSA.2023.0140336.

[106] Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. “Comparison of Deep Learning Models for Automatic Detection of Sarcasm Context on the MUStARD Dataset”. In: *Electronics* 12.3 (Jan. 2023), p. 666. ISSN: 2079-9292. DOI: 10.3390/electronics12030666. (Visited on 01/07/2025).

[107] Hamed Yaghoobian, Hamid R. Arabnia, and Khaled Rasheed. *Sarcasm Detection: A Comparative Study*. July 2021. DOI: 10.48550/arXiv.2107.02276. arXiv: 2107.02276 [cs]. (Visited on 01/09/2025).

Appendix

Table 11. Analyzed literature works

Publication type	Venue	Selected Studies
Journal article	ACM Trans. Asian Low-Resour. Lang. Inf. Process.	[37]
	Applied Sciences	[53], [54]
	ArXiv	[55]
	Argument & Computation	[56]
	Cognitive Science	[38]
	Computational Linguistics	[57]
	Computers	[58]
	Electronics	[40]
	Expert Systems	[59]
	Frontiers in Artificial Intelligence	[60]
	Heliyon	[61], [42]
	IEEE Access	[62], [63]
	Information Discovery and Delivery	[64]
	Information Processing & Management	[65], [66]
	International Journal of Advanced Computer Science and Applications	[67]
	International Journal of Advanced Computer Science and Applications (IJACSA)	[68], [33]
	International Journal of Computational Linguistics (IJCL)	[69]
	International Journal of Interactive Multimedia and Artificial Intelligence	[70]
	International Journal of Recent Technology and Engineering (IJRTE)	[71]
	International Journal on Recent and Innovation Trends in Computing and Communication	[72]
	Journal of Biomedical Informatics	[73]
	Journal of Computer Science	[74]
	Journal of Intelligent Systems	[46]
	Journal of Universal Computer Science	[75]
	Knowledge-Based Systems	[44]
	Language Resources and Evaluation	[76]
	Malaysian Journal of Computer Science	[77]
	Measurement: Sensors	[78]
	Multimedia Tools and Applications	[35]
	Neural Computing and Applications	[79]
	Pattern Recognition Letters	[80]
	Procedia Computer Science	[81], [32], [82]
	SN Applied Sciences	[83]
	Technical Reports on Language Technology	[84]
	Training, Language and Culture	[85]
Conference proceedings	2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)	[86]
	2021 International Joint Conference on Neural Networks (IJCNN)	[39]
	Advances in Information Retrieval	[87]
	CMNA@COMMA	[88]
	IberLEF@SEPLN	[89]
	International Conference on Computational Linguistics and Intelligent Systems	[34], [90]
	Natural Language Processing and Information Systems	[91]
	Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI	[92]
	Proceedings of the 12th Global Wordnet Conference	[93]

Continued on next page

Publication type	Venue	Selected Studies
	Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing	[94]
	Proceedings of the 28th International Conference on Computational Linguistics	[95]
	Proceedings of the 29th International Conference on Computational Linguistics	[45]
	Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)	[96], [97]
	Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature	[98]
	Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)	[99]
	Proceedings of the ACM Web Conference 2024	[100]
	Proceedings of the International Joint Conference on Neural Networks	[43]
	Proceedings of the Second Workshop on Figurative Language Processing	[41]
	Proceedings of the Twelfth Language Resources and Evaluation Conference	[101]
	Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019	[102], [36], [103]
Thesis	University of Passau	[104]

Table 12. Summary of survey articles on figurative speech detection.

ID	Figures of Speech Covered	Languages Covered	Time Span	Number of Studies	Main Insights
[51]	Antithesis, epanalepsis, zeugma, hyperbole, meiosis, irony, sarcasm, metaphor, rhetorical question, litotes, oxymoron, polysyndeton, chiasmus, and more	English (74 papers), German (10), Russian (8), French (4), Latin (4), Chinese (3), Czech (2), Japanese (1)	2006 - 2024	86 different approaches were analyzed across 39 primary studies	Focuses on lesser-known rhetorical figures, their definitions, datasets, and detection techniques. Highlights challenges such as dataset scarcity, rule-based approach reliance, and the need for deep learning methods. Suggests potential for improvements using LLMs.
[105]	Sarcasm, irony	Studies primarily focus on English	2019 - 2022	30 studies were reviewed	The review identifies deep learning as the most frequently used technique for sarcasm detection in recent studies. It emphasizes Twitter as the most commonly used data source and F1-score as the preferred evaluation metric. The study highlights the challenges of sarcasm detection due to ambiguity and context dependence, and suggests future directions such as exploring multimodal approaches and multilingual sarcasm detection
[106]	Sarcasm, irony.	English.	2010 - 2022	60 papers.	Model advancements from machine learning to deep learning and transformers; data selection remains a challenge; context incorporation improves performance.

Continued on next page

ID	Figures of Speech Covered	Languages Covered	Time Span	Number of Studies	Main Insights
[107]	Sarcasm and irony	English is the primary focus; other languages like Italian, Czech, Dutch, Greek, Indonesian, Chinese, and Hindi are also mentioned	2006 - 2021	87 papers were reviewed	The study identifies three major paradigm shifts in sarcasm detection: 1) semi-supervised pattern extraction, 2) hashtag-based supervision, and 3) incorporation of extra-textual context. The review emphasizes challenges such as sarcasm's implicit nature, the need for context-aware models, and the effectiveness of deep learning techniques
[52]	Sarcasm, irony, metaphor, simile, hyperbole, humor, satire	English (majority), with mentions of other languages such as Spanish and Arabic	2005 - 2019	120 studies analyzed	The review covers computational techniques for detecting figurative language on social networks, emphasizing challenges posed by non-literal text and the role of contextual understanding. It discusses feature extraction methods, dataset characteristics, and evaluation metrics used in different studies