A Survey of Figurative Language and Its Computational Detection in Online Social Networks

MUHAMMAD ABULAISH, South Asian University, India
ASHRAF KAMAL, Jamia Millia Islamia, India
MOHAMMED J. ZAKI, Rensselaer Polytechnic Institute, India

The frequent usage of figurative language on online social networks, especially on Twitter, has the potential to mislead traditional sentiment analysis and recommender systems. Due to the extensive use of slangs, bashes, flames, and non-literal texts, tweets are a great source of figurative language, such as sarcasm, irony, metaphor, simile, hyperbole, humor, and satire. Starting with a brief introduction of figurative language and its various categories, this article presents an in-depth survey of the state-of-the-art techniques for computational detection of seven different figurative language categories, mainly on Twitter. For each figurative language category, we present details about the characterizing features, datasets, and state-of-the-art computational detection approaches. Finally, we discuss open challenges and future directions of research for each figurative language category.

CCS Concepts: • **Networks** → **Online social networks**; • **Information systems** → *Data analytics*; • **Computing methodologies** → *Lexical semantics*; • **General and reference** → Surveys and overviews;

Additional Key Words and Phrases: Social network analysis, figurative language, sarcasm detection, irony detection, simile detection, metaphor detection, satire detection, hyperbole detection, humor recognition

ACM Reference format:
Muhammad Abulaish, Ashraf Kamal, and Mohammed J. Zaki. 2020. A Survey of Figurative Language and Its Computational Detection in Online Social Networks. *ACM Trans. Web* 14, 1, Article 3 (January 2020), 52 pages.
https://doi.org/10.1145/3375547

1 INTRODUCTION

The evolution of the Internet has led to its becoming a massive platform for human communication. Due to its wide reach, availability, and usefulness, it connects the world into a single meeting and information-sharing venue. The sharp rise of Web 2.0 changed the perception of the use of the Internet. Earlier, users were passively involved in Web 1.0, but there is a high degree of user involvement in Web 2.0. As a result, huge amounts of user generated content (UGC) is accessible

This article is an outcome of the R&D work undertaken project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation.
Authors' addresses: M. Abulaish, South Asian University, New Delhi, Delhi, India, 110021; email: abulaish@sau.ac.in; A. Kamal, Jamia Millia Islamia, New Delhi, Delhi, India, 110025; email: ashrafkamal.mca@gmail.com; M. J. Zaki, Rensselaer Polytechnic Institute, Troy, NY, New York, 12180; email: zaki@cs.rpi.edu.

1559-1131/2020/01-ART3 $15.00
https://doi.org/10.1145/3375547

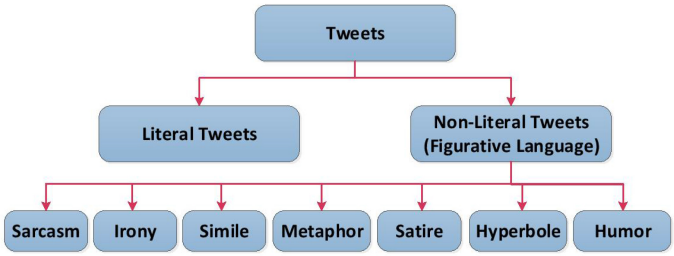

Fig. 1. A broad categorization of tweets.

world-wide [102]. Twitter[1] has emerged as a popular social networking service due to its micro-blogging nature. It allows users to post short messages of at most 280 characters, called *tweets*. As of the first quarter of 2019, Twitter's monthly active users averaged at 330 million.[2] The data available on Twitter are useful for varied purposes, such as product marketing, event monitoring, disease surveillance, trend analysis, election campaigns, e-governance, sentiment analysis, and open-source intelligence.

As shown in Figure 1, tweets can be categorized as *literal* tweets and *non-literal* or *figurative language* (FL) tweets. Literal tweets generally contain standard dictionary words and their underlying sentiment polarity is easy to determine, whereas non-literal tweets include words or phrases that reflect vivid and rhetoric effect, misleading the recognition of real sentiments expressed by the users due to the presence of figurative language. Depending on the nature of tweets, FL can be categorized as (i) *sarcasm*, (ii) *irony*, (iii) *simile*, (iv) *metaphor*, (v) *satire*, (vi) *hyperbole*, and (vii) *humor*.

Due to the uncertain sentiment polarity of figurative language, its computational detection is a non-trivial task and requires more research at the intersection of natural language processing, information extraction, and machine learning. Though figurative language can be found in any source of text, researchers have mostly concentrated on the detection of figurative language in Twitter, especially targeting *sarcasm* and *irony*, because of their contrasting nature within the tweets. However, some researchers have also considered the study of other figurative language categories, such as *metaphor*, *simile*, *satire*, *humor*, and *hyperbole*. Nevertheless, most of the existing surveys focus only on a few categories of figurative language. For example, a survey including different datasets, approaches, and trends is presented in Reference [85], but it is restricted to *sarcasm* only. Similarly, the survey in Reference [214] focuses only on the computational detection of *irony*.

To the best of our knowledge, there exists no comprehensive survey covering each category of figurative language in Twitter. Starting with a brief introduction to figurative language, its various categories, and a brief history of its evolution, we present an in-depth survey on the computational detection techniques for different figurative language categories. Our comprehensive survey covers literature on the detection of different categories of figurative language published between 2005 and 2019. The articles are sampled manually based on their publication venue and Google Scholar's citation statistics. Furthermore, we also present insights about the datasets, feature extraction techniques, evaluation metrics, validation approaches, and detection techniques for figurative language. At the end, we present a detailed discussion highlighting different open challenges and future directions of research in figurative language detection.

[1] https://www.twitter.com/.
[2] https://goo.gl/7Ermpu (last accessed on 20-Nov-2019).

Table 1. Exemplar Messages of Each Figurative Language Category

S. No.	Example	Source	FL Category
1	"I hate Australia in cricket, because they always win #sarcasm"	Bharti et al. [20]	Sarcasm
2	"Absolutely adore it when my bus is late #sarcasm"	Riloff et al. [181]	
3	"I'd like to thank Michele Obama for making the fruit snacks in the lunch room 90% tinier! Really changed my whole life with that one"	Mukherjee and Bala [142]	
4	"Plastic company making an ad on water pollution #irony"	Tweet (https://bit.ly/2kYEUXo)	Irony
5	"Gentlemen, you can't fight in here! This is the War Room #irony"	Filatova [57]	
6	"I don't want to be average. Is such as average thought"	Khokhlova et al. [100]	
7	"New law makes it legal for atheist doctors and nurses to refuse care to religious patients"	The science post (https://bit.ly/2tF6ElN)	Satire
8	"Holi: a Hindu festival of colors and secular festival of saving water"	Ravi and Ravi [171]	
9	"I'm extremely disappointed. Not as expected! It's just amazing how the flash works!"	Reganti et al. [175]	
10	"What do you use to talk to an elephant? An elly-phone"	Mihalcea and Strapparava [129]	Humor
11	"Infants don't enjoy infancy like adults do adultery"	Mihalcea and Strapparava [129]	
12	"me and my parents are so like-minded. Whatever I like, they mind"	Tweet (https://bit.ly/2mjrDJm)	
13	"Telling a teacher how to do their job because you have kids is like telling a dentist how to drill a cavity because you have teeth"	Tweet (https://bit.ly/2m93ogB)	Simile
14	"Jane swims like a dolphin"	Qadir et al. [166]	
15	"my neighbor is as cunning as a fox"	Hao and Veale [69]	
16	"He is the pointing gun, we are the bullets of his desire"	Jang et al. [82]	Metaphor
17	"My car drinks gasoline"	Shutova et al. [194]	
18	"Inflation has eaten up all my savings"	Shutova et al. [194]	
19	"Just tried yoga for the first time. I've never been more pissed off in my life"	Tweet (https://bit.ly/2muk8PH)	Hyperbole
20	"This is the best pizza in history"	Troiano et al. [207]	
21	"I won't wait for you: it took you centuries to get dressed"	Troiano et al. [207]	

2 FIGURATIVE LANGUAGE: BASIC DEFINITIONS

This section presents formal definitions of the figurative language categories. According to Hepburn [75], *figurative language* or *figure of speech* refers to "derivations from the strictly grammatical and logical modes of expression, by means of which ideas and thoughts are conveyed with vividness and force." Some of the commonly used figurative language categories found in online social media, especially in Twitter, are *sarcasm*, *irony*, *simile*, *metaphor*, *satire*, *humor,* and *hyperbole* [10, 14, 20, 61, 85, 166]. A formal definition of each figurative language category is given in the following paragraphs, and examples from each category are illustrated in Table 1.

Definition 2.1 (Sarcasm). The online Cambridge English dictionary[3] defines *sarcasm* as "*the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way.*"

The first three entries of Table 1 present three *sarcasm* examples. In the first example, *sarcasm* is expressed using a contradiction between the negative sentiment word "hate" and the positive situation phrase "they always win." In the second example, *sarcasm* is expressed using a contradiction between the positive sentiment word "adore" and the negative situation phrase "bus is late."

[3]https://dictionary.cambridge.org/dictionary/english/sarcasm.

Finally, the third example reflects *sarcasm*, since it mocks the fight against obesity project started by Michelle Obama, the former first lady of the USA.

Definition 2.2 (Irony). According to Hepburn [75], "*irony is a figure in which the literal import of the words is the contrary of what is meant to be expressed.*"

The entries 4 to 6 in Table 1 refer to the *irony* category. Example 4 reflects *irony*, which is clearly intended to criticize the plastic companies, reflecting the fact that plastic items are one of the main sources of water pollution. In example 5, *irony* is expressed by combining two sentences, but separately these sentences are non-ironic. Finally, the *irony* in example 6 is constructed using the negation phrase "I don't."

Definition 2.3 (Satire). The online Cambridge English dictionary[4] defines *satire* as "*a way of criticizing people or ideas in a humorous way, or a piece of writing or play that uses this style.*"

The entries 7 to 9 in Table 1 refer to the *satire* category. Example 7 is a *satire* post, since it uses satire in a humorous way to criticize doctors and nurses who refuse to treat LGBT patients citing a violation of their religious beliefs. In example 8, the first part of the sentence conveys literal meaning, while the second part "secular festival of saving water" criticizes existing customs, using *satire*. It refers to the opposite of the reality, since water is extensively used in "Holi" celebrations. Finally, in example 9, *satire* is expressed using reversal, in which the opposite of an actual situation is conveyed.

Definition 2.4 (Humor). The online Cambridge English dictionary[5] defines *humor* as "*the ability to be amused by something seen, heard, or thought about, sometimes causing you to smile or laugh, or the quality in something that causes such amusement.*"

The entries 10 to 12 in Table 1 refer to the *humor* category. Examples 10 and 11 induce *humor* by phonological ambiguity (elly-phone vs. telephone). It contains phonological information to generate *humor* along with a pun. As a result, it generates a funny result. Example 12 indicates that the person is amusing his/her parents using *humor*.

Definition 2.5 (Simile). According to Hepburn [75], "*simile is the explicit statement of the resemblance between two objects or notions belonging to different classes.*"

The entries 13 to 15 in Table 1 refer to the *simile* category. Example 13 conveys *simile* in which two different phrases that are semantically different from each other are compared and linked using the connecting word "like." Similarly in example 14, two unlike things, "Jane" and "dolphin," are connected to each other using "like," where Jane's swimming ability is compared with a dolphin. Finally, in example 15, a person's cunning nature is compared with a fox.

Definition 2.6 (Metaphor). According to Hepburn [75], "*metaphor is a trope founded upon resemblance. It is the substitution of one notion for another in virtue of some resemblance or analogy between them.*"

The entries 16 to 18 in Table 1 refer to the *metaphor* category. Example 16 clearly intends to compare people metaphorically with guns and bullets. In Example 17 and 18, the metaphorical usage of the word "drink" and "eaten up" join sentences of different concepts.

Definition 2.7 (Hyperbole). According to Hepburn [75], "*hyperbole consists in magnifying an object beyond the bounds of what is actual or even possible.*"

[4] https://dictionary.cambridge.org/dictionary/english/satire.
[5] https://dictionary.cambridge.org/dictionary/english/humor.

The entries 19 to 21 in Table 1 refer to the *hyperbole* category. Example 19 conveys *hyperbole*, which clearly puts exaggeration through the phrase "pissed off" to create emphasis in the text. Example 20 expresses the qualitative aspect of *hyperbole* using the word "best." Finally, example 21 expresses *hyperbole* by referring and joining concepts with different intensities and emphasis.

3 A BRIEF HISTORY OF FIGURATIVE LANGUAGE STUDIES

The study of figurative language goes back a few decades, and it is one of the well-studied topics in interdisciplinary sciences, such as philosophy, psychology, linguistics, cognitive science, and neuroscience. Leggitt and Gibbs [110] mention that negative sentiments such as anger, frustration, and hatred are mainly expressed using figurative language.

Sarcasm is one of the main categories of figurative language. Wilson [219] suggests that sarcasm occurs in texts and contextual information due to their situational imbalance. Giora [64] states that negation is implicitly involved in sarcasm or irony without explicitly mentioning any negation word, such as *not*. Further, Bowes and Katz [25] consider sarcasm and irony as a means to attack any particular target. McDonald [126] describes sarcasm as "intimately associated with particular negative affective states" (where "affective" relates to moods, feelings, and attitudes). According to Wilson and Sperber [220], verbal irony should be considered as echoic to maintain the difference in use and mention. The term "echoic" means that a speaker "tactically dissociates herself from an attributed utterance or thought." Sperber and Wilson [195] consider irony as an echoic allusion to a thought or utterance. Utsumi [211] classifies an expression as ironic when it involves a situation in an ironic surrounding if it covers three conditions, namely,

- "The speaker has an expectation E at temporal location t_0."
- "The speaker's expectation E fails at temporal location t_1."
- "As a result, the speaker has a negative emotional attitude toward the incongruity between what is expected and what actually is."

As seen above, irony often occurs along with sarcasm, but it is a major category of figurative language in its own right. Expression of irony using the opposite question mark symbol was first noticed by a French poet, Alcanter de Brahm, to help readers in understanding the presence of irony in an utterance [190]. The pragmatic theory proposed by Grice [65] assumes that an ironic utterance is considered by a hearer when he/she receives an alert of violation of pragmatic principles, such as *maxim of quality* in which "speakers should not say what they believe to be false." However, this theory could not survive, as it failed to explain many ironic utterances. Later on, Clark and Gerrig [36] proposed the *pretense* theory of irony in which the speaker pretends to be an injudicious person speaking to an audience, and the speaker aims the irony-addressed person to recognize the pretense and thereby his/her attitude towards the utterance, the audience, and the speaker. Later on, in line with the *pretense* theory, Kumon-Nakamura et al. [105] proposed the *allusional pretense* theory, wherein an ironic utterance is not only "pragmatically insincere" but also alludes to a "failed expectation." In References [38, 144], the authors argued that generally children aged 5 to 6 years old are good candidates for irony recognition. The children in this age group can construct sarcastic or ironic utterances in a better manner as compared to adults. A brief study on the use of irony and sarcasm is available in linguistics and psychological sciences [91, 141, 173].

The presence of irony is generally also found in satire, and if the audience does not get the actual sense with respect to the ironic dimension, then satire loses its importance [9]. As discussed in Frye [60], the acceptance of satirical messages among the writers and readers is based on a common agreement. According to the Highet [78], satire must have an aim "to cure folly and to punish evil." Condren [37] stated that satire is in the form of *Juvenalian* or *Horatian* styles. The Juvenalian

style of satire is based on ridicule and sarcasm, whereas the Horatian style contains tease and humor. Though various studies on satire are found in literature [101, 107, 122, 160], computational approaches are rare.

Humor is well-studied in areas such as philosophy, linguistics, psychology, and cognitive science [130, 227]. Accordingly, there are several theories based on it, such as superiority [66], release [185], and incongruity [169]. In addition, some linguistic theories based on humor, such as general theory of verbal humor [7] and ontological semantic theory of humor [170], are also available. The relatedness of funny content in humor varies on the basis of culture, language, and region. For example, jokes that induce laughter in theaters in India hardly put a smile on a Dutch person [44]. Attardo [6] points out that verbal humor is linked with knowledge resources, such as language, situation, and opposition, to create funny effect. Ruch [184] explained the linkage between personality and appreciation of humor. Hertzler [77] considered sociological aspects (i.e., cultural patterns) to categorize humor. The usage of humor also increases interpersonal attraction in people [30]. Hay [70] described humor categories, such as *wordplay*, *fantasy*, *insult*, *jokes*, *self-deprecation*, and *vulgarity*.

In simile, dissimilar entities are compared using connecting words, such as *like*, *as*, and *than*. Israel et al. [80] state that comparing fundamentally different types of entities is what makes a simile figurative. A simile can be both open and closed [17]. Qadir et al. [167] found that 92% of similes are open. In open simile, the shared property between two entities is implicitly involved. Consider the example, "My room feels like Antarctica" taken from Reference [167]. In this example the word "cold" acts as a shared property that is implicitly involved. The authors also mentioned closed simile in which the shared property between two entities is explicitly involved. Consider the example, "My room is as cold as Antarctica" taken from Reference [167], where the word "cold" acts as a shared property that is explicitly involved. According to Hanks [68], simile vehicles related to semantic categories (e.g., animal, artifact) are generally common. Niculae and Danescu-Niculescu-Mizil [148] identified the following constituents of a simile for its characterizations:

- *Topic or tenor*: Act as subject comparison indicator.
- *Vehicle*: Object that is used for comparison.
- *Comparator*: The connecting words such as "*as*," "*like*," or "*than*."
- *Event*: Act/state.

They also proposed an optional component *property*, which indicates a shared attribute. Consider the example "sterling is much cheaper than gold," taken from Reference [148]. In this example, the words "sterling" and "gold" are considered as *topic* and *vehicle*, respectively, whereas the words "is," "cheaper," and "than" are considered as *event*, *property*, and *comparator*, respectively.

Unlike simile, metaphor aims to compare two dissimilar entities without using any connecting words. Researchers from various disciplines, such as psychology, linguistics, sociology, anthropology, and computational linguistics have studied the problem of metaphor [136]. In References [124, 194], they collected metaphoric expressions from a manually annotated seed set. Further, using these seed metaphoric expressions, their system yields a large number of similar metaphoric structures from the corpus. Martin [123] implemented a metaphor database called Metaphor Interpretation, Denotation, and Acquisition System to search for metaphors encountered in text documents.

Hyperbole[6] differs from comparison-based figurative language categories such as metaphor and simile due to the presence of overstatement to reflect humorous effects. In fact, hyperbole is often used in satire, sarcasm, humor, and irony. Kreuz and Roberts [104] reported that hyperbole

[6] https://literarydevices.net/hyperbole/.

and ironic tone of voice can be considered jointly to detect verbal irony. Kreuz and Caucci [103] found that the usage of hyperbole indicators such as interjections and intensifiers are the clues for sarcastic texts.

4 FEATURE EXTRACTION TECHNIQUES

Having briefly considered the inter-disciplinary history of FL studies, this section presents a description of different feature extraction techniques in the study of figurative language categories.

4.1 Feature Extraction Techniques for Sarcasm Detection

Various types of features have been for sarcasm detection, in supervised, semi-supervised, rule-based, and linguistics-based approaches [1, 8, 15, 21, 23, 24, 42, 106, 111, 112, 117, 145, 168, 181, 198, 208, 210, 217]. Below, we list the feature types and their descriptions.

- *Pattern-based features* [20, 23, 87, 111, 119, 181, 208]: Pattern-based features are used to determine sarcasm in text messages. Riloff et al. [181] considered "positive sentiment verbs and negative situation phrases" for sarcasm detection in Twitter. Consider the tweet, "Oh how I love being ignored #sarcasm," taken from Reference [181]. A bootstrapped learning method is applied to collect positive sentiment verbs, i.e., "love" and phrases of negative situations, i.e., "being ignored." Similarly, Bharti et al. [20] proposed an algorithm "interjection-word start" for pattern-based feature extraction. Consider the example, "Wow, what an amazing night this has turned out to be #sarcasm," taken from Reference [20].

  A pattern encoding the presence of an interjection, i.e., "wow," followed by the presence of an intensifier (adjective or adverb), i.e., "amazing," can be used to construct a feature.
- *Hyperbolic features* [20, 21, 106]: Hyperbolic features are used to indicate over-statement or exaggeration in text, which adds extra emphasis in sarcasm-related utterances. The frequent usage of adjectives or adverbs is a key indicator of hyperbole. Hyperbolic features are constructed using NLP tools, such as spacy,[7] NLTK,[8] and Stanford Parser,[9] to identify over-statement Parts-Of-Speech (POS) tags. Consider the example, "fantastic weather when it rains," taken from Reference [111], in which "fantastic" is an adverb/adjective, whereas "wow" is an interjection.
- *Syntactic features* [23, 24, 87, 111, 112, 145]: Syntactic features are the most commonly used features for sarcasm detection. They include presence of interjections, bag-of-words (n-grams), capitalized words, stopwords, POS tags (e.g., adverb, pronoun, adjective, and verbs), negations, and text lengths. Consider the example, "wow I love it WHEN I am called at 4 a.m. because my neighbour's kid can't sleep!" taken from Reference [24]. In this example, syntactic features include POS tags, such as, *"I" –> "PRP"* and *"Love" –> "VBP,"* where PRP stands for personal pronoun and VBP represents verb, non-third person singular present. For stopwords like "am, at" and interjections like "wow," count is the number of their occurrences in the text. Bag-of-words include n-grams such as "(wow, I)" and "(wow, I, love)." For accurate tagging, negations like "can't" are replaced by "can not." This is done via a contraction list, which is a dictionary in which every negation word is a key, along with its full-form value.
- *Sentiment-based features* [1, 8, 15, 23, 24, 111, 112, 117, 145, 168, 181, 198, 217]: Sentiment-related features are used to deal with the polarity of sarcastic utterances. They include features based on positive and negative words, emotional words, positive and negative phrases,

[7] https://spacy.io/.
[8] https://www.nltk.org/.
[9] https://nlp.stanford.edu/software/tagger.shtml/.

and sentiment score. Consider the example, "oh how I love being ignored," taken from Reference [181]. Here, "love" is a positive word, "ignored" is a negative word, and "being ignored" is a negative phrase. Sentiment lexicons, such as Linguistic Inquiry and Word Count (LIWC),[10] AFINN[11] (an affective lexicon by Finn Årup Nielsen), SenticNet,[12] and Sentisense[13] have been used by many researchers to construct sentiment-based features.

- *Pragmatics features* [23, 24, 61, 87, 117, 134, 143, 145]: Pragmatics features use counts and the presence of elements such as smileys, emoticons, reply, and @user that are generally embedded within the texts. Consider the example, "@UserName that's what I love about Miami. Attention to detail in preserving historic landmarks of the past," taken from Reference [145]. Here, the constituent @UserName can be used as a pragmatic feature. Emotion-related lexicons, such as EmoLex[14] and EmoSN,[15] are generally used to construct pragmatics features.
- *Punctuation-based features* [21, 24, 42, 106, 145, 165, 208, 210]: Punctuation-based features are represented as exclamation marks, question marks, and quotes. The extra presence of such markers within a text indicates the presence of sarcasm. Consider the example, "all your products are incredibly amazing!!!" taken from Reference [24]. Here, the excessive use of exclamation marks is a punctuation-based feature, which can be constructed by counting its occurrences.
- *Linguistic features* [87, 92, 134, 143, 145]: Linguistic features are also known as lexical features. They are represented as implicit and explicit incongruities, intensifiers, exclamation marks, adverbs, and adjectives. Implicit incongruity is represented using implied sentiment phrases. Consider the example, "I love this paper so much that I made a doggy bag out of it," taken from Reference [87]. Here, the phrase "I made a doggy bag out of it" contains implied sentiment and the polarity word "love" is incongruous with the implied sentiment. However, explicit incongruity is represented using both positive and negative polarity words. Consider another example, "oh how I love being ignored," taken from Reference [181], in which the positive word "love" and negative word "ignored" are used.
- *Self-deprecating features* [2]: Self-deprecation can be found in *self-around* instances, defined as cases where users text about themselves. These consist of patterns such as "*interjection* followed by token *I*," "token *I* followed by *verb* and *question word*," "common deprecating patterns," "token *I* followed by *verb* and *adverb* or *adjective*," "token *am* followed by *adjective* or *adverb*," and "token *I* followed by *negative modal verb*." These features are mainly extracted using POS tags and patterns associated with self-around keywords, such as "I," "my," and "me." Consider the example, "I love being ignored; it feels good. #bigleague #sarcasm," taken from Reference [2], in which the phrase "love being ignored" is referred as self-deprecating sarcasm.
- *Twitter-specific features* [8, 87, 125, 168]: Some authors also consider Twitter-specific metadata-based features, such as the author's historical topics, profile information, historical salient terms, profile unigrams, and author historical salient terms.
- *Other features*: Gaze-related features based on eye-tracking of the annotators have been used in Joshi et al. [88] for modeling sarcasm understandability. Readability features are used in Rajadesingan et al. [168] to measure tweet complexity in terms of expression. It consists

[10] http://liwc.wpengine.com/.
[11] https://goo.gl/yEiQmG.
[12] https://sentic.net/.
[13] http://nlp.uned.es/~jcalbornoz/SentiSense.html.
[14] https://www.saifmohammad.com/WebPages/lexicons.html.
[15] https://www.gelbukh.com/emosenticnet/.

of features such as number of words, syllables, polysyllables (i.e., more than one syllable), syllables per word [59], and polysyllables per word [108]. Joshi et al. [90] also consider word embedding–based features for sarcasm detection.

4.2 Feature Extraction Techniques for Irony Detection

This section presents a brief description of different feature categories for irony detection, which have been used mainly in supervised machine learning approaches [12, 13, 22, 72, 96, 97, 172, 179, 180].

- *Frequency-based features* [12–14]: These features are used to capture frequency imbalance between words, such as finding a gap between the rare words and common words. For example, consider the example tweet,[16] "CHANDLER: I am so glad we are having this *rehearsal* dinner. You know, I rarely get to practice my meals before I eat them," in which *rehearsal* is the rare word. These features have been constructed using the ANC[17] frequency data corpus.
- *Written-spoken features* [12–14]: Tweets are presented in written forms in which usually spoken styles are employed by the users; that is, a word may be used both in written and spoken style due to the informal style of writing in tweets. These features are also constructed using the ANC data corpus.
- *Signature-based features* [180]: These features represent textual markers or signatures in ironic utterances. Signature-based features are used to highlight certain aspects in a text, using capital words and quotes. *Pointedness*, *counterfactuality*, and *temporal compression* are three dimensions in signature-based features. Pointedness indicates explicit marks, such as ?, :, ;, and !. Counterfactuality indicates implicit marks through usage of discursive terms such as "about," "yet," and "nonetheless." Finally, temporal compression focuses on elements that indicate opposition in time, such as temporal adverbs like "suddenly," "now," and "abruptly." Consider the example, "I HATE to admit it but, I LOVE admitting things *!!*" taken from Reference [180]. Here, the usage of capitalized words such as "I HATE" and "I LOVE" highlight signature-based feature.
- *Unexpectedness features* [179, 180]: Unexpectedness and incongruity are used as an indicator for irony [118]. Unexpectedness features are used in ironic texts to represent temporal and contextual imbalances. Consider the example, "I hate that when you get a girlfriend most of the girls that *didn't want you* all of a *sudden want you!*" taken from Reference [180]. Here, the temporal imbalance is related to the degree of opposition as compared to the information described in the present and past tenses (e.g., "didn't want you" and "sudden want you"), whereas contextual imbalance is used to capture inconsistencies in the context. Unexpectedness features are constructed using the Resnik measure [158], which calculates pairwise semantic similarity from WordNet [132].
- *Style-based features* [179, 180]: These include character-grams, skip-grams, and polarity skip-grams. Consider the example, "*there* are far *too* many crazy people in my psychology class *exactly*," taken from Reference [180]. Character-grams consider sequences of morphological information, i.e., affixes and suffixes (e.g., ly). Skip-grams consider gaps between words, such as "there, too." However, polarity skip-grams consider abstract sequences of text based on polarity of positive and negative terms rather than specific content words (characters). The main assumption behind this feature is that usually in ironic sentences

[16] https://twitter.com/friendsreruns/status/714803445493010432?lang=en.
[17] https://www.anc.org.

positive words are taken to convey a negative meaning. To construct these features, the Macquarie Semantic Orientation Lexicon (MSOL) [135] is applied. Consider the example, "I need more than luck. I need Jesus and I'm an atheist . . . ," taken from Reference [180]. Here, using MSOL and applying two-word skips after stopwords removal, an abstract representation can be obtained from sequences of positive and negative polarity label tags, i.e., pos_{need} $\text{pos}_{\text{jesus}}$ $\text{neg}_{\text{atheist}}$.

- *Emotional scenario features* [179, 180]: In textual content, emoticons are used to convey information such as mood, feelings, and sentiments. In ironic texts, emotions provide a platform for any situation to be ironic. Consider the example, "I feel so miserable without you, it's almost like having you here," taken from Reference [180], which is ironic in nature. Emotional scenario features aim to capture emotion in the form of mood, sentiments, and feelings to convey irony in favorable and unfavorable contexts. Emotional scenario features span over three dimensions, namely *activation*, *imagery*, and *pleasantness*. Activation refers to the degree of response as passive or active, usually shown by humans in an emotional scenario. Imagery refers to the way of dealing with a mental picture for a given word. Finally, pleasantness refers to the degree of pleasure suggested by a word. To detect these emotion-based dimensions, the Whissell's dictionary of affect in language [218] is used, which contains 8K English words and scores for the different dimensions.
- *Polarity features* [179]: These include words that indicate either positive or negative semantic orientation. Consider the example, "it was so cold last winter that I saw a *lawyer with his hands in his own pockets*," taken from Reference [179]. This example reflects negative semantics towards the lawyer via the phrase "lawyer with his hands in his own pockets." To construct polarity features, the MSOL lexicon [135] is applied, which contains 30,458 positive entries and 45,942 negative entries.
- *Surface features* [97]: These include tweet length, presence and absence of punctuation, interjections, emoticons, and slang words. Consider the example, "wow that's a huge discount, I'm not buying anything *!!!*" taken from Reference [20]. Roze et al. [182] also used a French lexicon to construct surface features for irony detection in French.
- *Shifter features* [97]: In this feature category, we check whether a tweet consists of an intensifier (e.g., adverb, adjective) and negation words or verbs. Consider the example, "*wow*, that's a huge discount, I'm *not* buying anything!!" taken from Reference [20], in which "not" is used as a shifter feature.
- *Sentiment shifter features* [97]: Words and expressions can affect the polarity of a text. These features determine whether a tweet contains an opinion word that lies under the scope of an intensifier adverb. Consider the example, "*effectively*, you did not do much at work today. *great*!" where the opinion word *great* is within the scope of the adverb *effectively*.
- *Opposition features* [97]: These features are inspired by the work of Riloff et al. [181] and are based on lexico-syntactic patterns. They check whether a tweet contains opposition in sentiment, or positive (negative) contrast between the subjective and objective propositions. Consider the example tweet,[18] "absolutely *love* it when *my bus is late*." Here, we can notice a contrast between the subjective proposition and the objective one.
- *Other features*: Psycho-linguistic features are considered in Reference [172]. These features are implemented using LIWC, which contains a psychological dictionary. Bosco et al. [22] proposed *polarity reversing* and *emotion expression* features. *Polarity reversing* aims to reverse the polarity of a positive expression as negative and vice versa. Consider the example, "we are on the cliff's edge, but with me we will make a great leap forward," taken from

[18] https://twitter.com/MagduhS/status/190247374864658432.

Reference [22], which uses polarity reversing features. However, *emotion expression* consists of emotion-related words, such as "anger," "love," "fear," "joy," and "sadness." Hee et al. [72] considered *contrasting evaluation* in which contrast can be examined using explicit and implicit evaluations, i.e., polarity can be judged using contextual clues or world knowledge. Consider the example, "I *cannot wait* to go to the dentist later!" taken from Reference [72]. In this example, though going to the dentist is an unpleasant situation, the phrase *cannot wait* indicates a positive evaluation contrasted by the act of going to the dentist, and indicating a negative sentiment.

4.3 Feature Extraction Techniques for Satire Detection

Satire detection features, listed below, have been used mainly for supervised learning [29, 175, 183].

- *Predictive features* [183]: These include features such as *absurdity*, *grammar*, *negative affect,* and *punctuations*. The *Absurdity* feature is the presence of unexpected entries, such as names of people, locations, and places in the final sentence of the satirical news. Consider the example, "at press time, researchers from Christopher Hitchens Memorial University discovered that it was fun to drink a lot of Johnny Walker Red Label and call people sheep," taken from the final line of the Canadian online satirical newspaper *The Beaverton's*.[19] To extract *absurdity* features, NLTK POS tagger and Named Entity (NE) recognizer[20] are used for named entities recognition. Grammar features refer to the count of POS tags, such as adjectives, adverbs, pronouns, conjunctions, and prepositions. *Negative affect and punctuation* features consider the presence of negative affect terms (extracted using the LIWC dictionary) and punctuations, such as question marks, exclamation, and quotes in the satirical text.
- *Headline features* [29]: Satirical news or articles can be recognized usually from their headline contents [29]. Headline features include the presence of headline tokens twice—first, token the news headline, and second, the same token from the news body. Consider the following example taken from satirical newspaper *The Onion*[21]:
  Headline: "God answers prayers of paralyzed little boy."
  News body: "While one God's response came at approximately 10 a.m. Monday, following a particularly fervent Sunday prayer session by little Timmy."
  Here, the token "God" is counted twice, first as a token in news headline, and second in the news body.
- *Profanity feature* [29]: Burfoot and Baldwin [29] mention that non-satirical news articles usually do not include profanity (offensive) language, but satirical news contains profane content, used as a humorous device to show exaggeration. Profanity feature is considered as a binary feature (obtained using, for example, `Regexp::Common::profanity` Perl module[22]). Consider another example, "*Black Guy* asks nation for change," taken from *The Onion*. Here, *Black Guy* is an offensive remark for the former US president, Barack Obama.
- *Slang feature* [29]: Similar to profanity, satirical articles also contain slang features. Consider the example,[23] "I'm talking about my friends *IRL* not you, loser." Here, IRL is slang that

[19] https://www.thebeaverton.com/2015/08/scientists-at-university-of-the-lord-discover-that-jesus-is-lord/.
[20] https://nlp.stanford.edu/software/CRF-NER.shtml.
[21] https://www.theonion.com/god-answers-prayers-of-paralyzed-little-boy-1819564974.
[22] https://metacpan.org/pod/Regexp::Common::profanity.
[23] https://www.urbandictionary.com/define.php?term=IRL.

stands for "in real life." This feature is constructed by checking the use of each word as a slang in Wiktionary.[24]

- *Sentiment amplifier feature* [175]: Satirical texts consist of sentiment amplifier features. These features highlight and intensify the emotional elements in text. Satirical texts indicate high emotions in the form of emoticons, acronyms, and interjections. Acronyms (e.g., "LOL") and emoticons (e.g., smiling face ":)" and sad face ":(") are used to generate such features.
- *Sensicons feature* [175]: Sensicons refer to the five type of senses—*sight*, *hearing*, *taste*, *smell*, and *touch* from sensorial lexicons [203]. Satirical texts consider these sensicon senses to show disgust and anger. The sensorial lexicons contain sense association score for these five senses, which are considered as individual features. Consider the example, "when the word *apple* is uttered, the average human mind will visualize the appearance of an apple, stimulating the eye-sight, feel the smell and taste of the apple, making use of the nose and tongue as senses," taken from Reference [175], which lists many sensicon features.

4.4 Feature Extraction Techniques for Humor Recognition

Features for humor recognition have been used mainly in supervised learning approaches [18, 115, 130, 174, 224, 225, 227]. We detail them below.

- *Incongruity features* [115, 224]: Humor indicates incongruity in the form of opposition or contradiction. Consider the example, "a clean desk is a sign of a cluttered desk drawer," taken from Reference [224], which presents incongruity and contrast using the phrases "clean desk" and "cluttered desk drawer." To construct this feature, word2vec [131] has been used to measure *disconnection* (i.e., maximum semantic distance of word-pairs) and *repetition* (i.e., minimum semantic distance of word-pairs) in a sentence.
- *Ambiguity features* [115, 224]: In this feature category, words with multiple meanings are considered for humor recognition. Consider the example, "did you hear about the guy whose whole left side was cut off? He's all right now," taken from Reference [224], which contains ambiguity features. These features include *sense combination*, *sense farmost*, and *sense closest*. To calculate *sense combination*, possible meanings of a word are determined and then all senses are aggregated as $\log(\prod_{i=1}^{k} n_{w_i})$, where n_{w_i} represents the total number of senses of word w_i. The lexical resource WordNet [54] is used to construct this feature. Similarly, for a given sentence, *sense farmost* and *sense closest* are used to calculate the largest and smallest path similarity[25] between the senses of a word, respectively.
- *Interpersonal effect features* [115, 224]: This feature category includes *sentiment* and *subjectivity* related features, such as count of negative or positive words, and count of weak or strong subjectivity–oriented words. Consider the example, "your village called. They want their Idiot back," taken from Reference [224]. Here, the word "idiot" shows strong sentiment that carries humor. Interpersonal effect features are generally constructed using TextBlob.[26]
- *Phonetic features* [115, 224, 227]: Humorous texts often contain incongruous sounds or words. The presence of phonetic properties in humorous texts is an important clue [130]. The presence of phonetic attributes generates comic effect and makes the texts humorous. *Alliteration chain* and *rhymes chain* are considered as phonetic features in automatic humor recognition tasks [224]. *Alliteration chain* refers to the beginning of two or more words with

[24] https://www.wiktionary.org/.
[25] http://www.nltk.org/howto/wordnet.html.
[26] https://textblob.readthedocs.io/en/dev/.

the same phones. Consider the example,[27] "Dan's dog dove deep in the dam, drinking dirty water as he dove." Here, *rhymes chain* refers to the relationship when two words end with the same syllable. Consider the example, "what is the difference between a nicely dressed man on a tricycle and a poorly dressed man on a bicycle?" taken from Reference [224]. Here, the number of alliteration/rhymes chains in a text, and the maximum length of alliteration/rhymes chains can be considered as features. Phonetic features are extracted using the CMU Pronouncing Dictionary.[28]

- *Stylistic features* [130]: This category includes *antonymy* and *adult slang* related features. Humor reflects the comic effect due to the presence of antonyms in a sentence. Consider the example, "always try to be *modest* and be *proud* of it!" taken from Reference [130], which contains stylistic features. WordNet [132] is used to capture antonyms in the sentence. Adult slang-based humor is very famous. Consider the example, "behind every great man is a great woman, and behind every great woman is some guy *staring at her behind*!" taken from Reference [130], where the phrase *staring at her behind* indicates an adult slang. To construct adult slang feature, WordNet Domains[29] has been used in conjunction with synsets[30] labeled for domain "SEXUALITY."
- *Homophones feature* [18, 174]: This feature includes words with the same pronunciation in a sentence, which can be recognized as a clue for humor. Consider the example, "what is everybody's favorite aspect of mathematics? Knot theory, that's for sure," taken from Reference [18], in which sound-alike words are "knot" and "not." The CMU Dictionary has been used to obtain homophones of words in a sentence.
- *Homographs feature* Beukel and Aroyo [18]: This feature includes words with two definitions in a sentence. Consider the example, "Cliford: The Postmaster General will be making the toast. Woody: wow, imagine a person like that helping out in the kitchen!" taken from References [18, 202]. Here, the word "toast" indicates multiple meanings. Lists of homographs available on Wikipedia[31] and WordNet are used to construct this feature.
- *Affective polarity feature* [225]: Humorous sentences make people laugh, which reflects emotion. This feature includes emotion polarity and intensity. Like subjectivity feature, the affective polarity score can also be calculated using TextBlob.

4.5 Feature Extraction Techniques for Simile Detection

As discussed in Section 3, simile is composed of four components—*tenor*, *vehicle*, *event*, and *comparator*. Consequently, features for simile detection are related to these components. This section presents a brief description of simile-specific features, which have generally been used in supervised machine learning approaches [166].

- *Lexical features*: In simile texts, lexical features include simile components, paired components, vehicle pre-modifier, and explicit properties. As discussed in Reference [166], simile components are binary features for *tenor*, *vehicle*, and *event* phrases. For example, the word "dog" as a *tenor* is different from the word "dog" as a *vehicle*. Pair components are considered as binary features in which a pair of components indicates affective polarity. For example, "*event*:feel, *vehicle*:ice box" indicate negative polarity for tenors such as, "house," "room," and "hotel." Vehicle pre-modifier is considered as a binary feature for every noun or

[27] http://waysoffigurativelanguage.weebly.com/alliteration.html.
[28] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
[29] http://wndomains.fbk.eu/.
[30] http://www.nltk.org/howto/wordnet.html.
[31] https://en.wikipedia.org/wiki/List_of_English_homographs.

adjective pre-modifier associated with the vehicle, such as "smells like wet rat" and "smells like wet shirt." These features are considered in Qadir et al. [166].

- *Semantic features*: These include hypernym class and perception verbs. Similar words, such as "room" and "bedroom" are considered in the same hypernym class, which is used in different similes with the same affective score. Hypernyms in simile texts are obtained with the help of WordNet [132]. Perception verbs are commonly seen in similes. Each verb is considered as a binary feature if the event component indicates the perception verb. Consider the example, "*looks* like a model," taken from Reference [166]. These features are considered in Qadir et al. [166].
- *Sentiment-based features*: In simile texts, semantic features include component sentiment, explicit property sentiment, sentiment classifier label, and simile connotation polarity. In component sentiment, *tenor*, *vehicle*, and *event* components are considered. A total of three binary features are constructed for each component to capture the presence of a positive sentiment word. Similarly, three binary features are used to capture the presence of negative sentiment words. Explicit property sentiment is considered as a numeric feature that is used to count the number of positive (negative) *properties* associated with vehicle. Sentiment lexicons such as AFINN [149] and Multi-Perspective Question Answering [221] have been used to capture the *property* words. Sentiment classifier label features are considered as a binary feature for positive and negative label representation as per the National Research Council (NRC)-Canada[32] sentiment classifier for simile assignment. Simile connotation polarity feature is used as a binary feature for positive and negative words using a connotation lexicon [55]. Consider the example, "acts like a *celebrity*" and "smells like *garbage*," taken from Reference [166], where *celebrity* and *garbage* indicate positive and negative connotation, respectively. These features are considered in Qadir et al. [166].

4.6 Feature Extraction Techniques for Metaphor Detection

This section presents a brief description of metaphor-specific features, which have generally been used in supervised machine learning approaches [81, 83]. Many of these features rely on *target words*, with usage in both literal and metaphoric sentences [81], and provide an important clue for metaphor detection.

- *Topic Transition–based features* Jang et al. [81]: Metaphoric words in sentences are incohesive with the context. It is important to consider semantic or topical cohesion for metaphor detection. Sentence Latent Dirichlet Allocation (LDA) determines if a target word resides in a sentence, and the topic changes around it. Using sentence LDA [84] topic transition–based features, such as *target sentence topic*, *topic difference*, *topic similarity*, *topic transition*, and *topic transition similarity,* are captured. Target sentence topic is a T-dimensional binary feature (where T is the number of topics), which indicates whether the topics in a sentence consist of the target word. Topic difference is designed with an assumption that metaphoric sentences are likely to be different from their neighboring sentences in terms of topics (i.e., left- and right-side sentences). This feature is a two-dimensional binary feature that indicates how much the target sentence topic differs from its neighboring sentences. Topic similarity is a two-dimensional feature that stores a value between 0 and 1. This feature captures the similarity between the topics of the target sentence and the sentences that are before and next to the target sentence. Topic transition is a $2 \times T$-dimensional binary feature that indicates the difference between the topics of the target sentence and its

[32] https://www.nrc-cnrc.gc.ca/index.html.

neighboring (previous and next) sentences. Topic transition similarity is a two-dimensional feature of continuous values. This feature captures the cosine similarity between the topics of the target sentence and its neighboring sentences.

- *Global contextual-based features* [83]: Features that span sentence boundary in a corpus are considered as global contextual features. Features including *semantic word category*, *topic distribution*, and *lexical chain* are examples of global contextual features.
- *Local contextual-based features* [83]: Features that are restricted to the sentence boundary within a corpus are considered as local contextual features. Features including *semantic relatedness*, *lexical concreteness*, and *grammatical dependency* are the examples of local contextual features. Semantic relatedness represents the semantic similarity between a pair of words, and it is calculated using the cosine similarity between the words, based on their topic distributions. If semantic relatedness between a target word and the context words is low, then the target word is considered as a metaphor. Lexical concreteness ensures that the lexical usage in a sentence follows the norms of the underlying language. It is measured using concreteness ratings database [28] and considered as an important clue to detect metaphors. Finally, grammatical dependency represents the asymmetrical relations called *dependencies* between the lexical elements of a sentence.

5 DATASETS

This section describes the datasets considered by various researchers for figurative language categories. A summary of these datasets and their characteristics is given in Table 2.

5.1 Sarcasm-related Datasets

We have broadly classified the datasets used in the existing studies into three categories: (i) Twitter datasets, (ii) long-text datasets, and (iii) others.

- *Twitter datasets*: Twitter is a commonly used platform for sarcasm and, accordingly, most of the researchers have considered crawled tweets (e.g., using the Twitter REST API or Streaming API Bharti et al. [21]). Though many researchers have crawled Twitter datasets for their studies, they are not allowed to publish them on the Web. As a result, some authors (e.g., References [61, 112, 165]) have posted only tweet IDs in the public domain, and the respective tweets and metadata can be fetched using the Twitter API. However, some of the authors (e.g., References [168, 181]) restrict data access further and provide tweet IDs only on request. There are also some websites (e.g., http://thesarcasmdetector.com/ and http://twiqs.nl/) that provide free access to Twitter datasets without any prior permission.

  Since most of the researchers have considered sarcasm detection as a binary classification problem for which annotated datasets are required, there have been different approaches to annotate Twitter datasets.

  — *Manually Annotated Tweets (MAT)*: In a manual annotation approach, a tweet is manually labeled as sarcasm or non-sarcasm based on human judgment. For example, in References [1, 125, 165, 181], their datasets were all manually annotated. Since manual annotation is a restrictive and time-consuming process, crowdsourcing platforms like Amazon Mechanical Turk[33] have also been used in some studies (e.g., Reference [42]) for manual annotations of tweets.

  — *Hashtag-annotated Tweets (HAT)*: In this approach, tweets are annotated based on the hashtags that are generally used as bookmarks or labels to express the real intent behind

[33] https://www.mturk.com/.

Table 2. A Summary of Datasets Used in Various Figurative Language Detection Studies

Category	Sources/Web	Online links	# Instances	Labeled?
Sarcasm, Irony, Metaphor	Ghosh et al. [61]	http://alt.qcri.org/semeval2015/task11/	9,000	Yes (HAT)
Sarcasm, Irony	Filatova [57]	Contact author	1,254	Yes (Manual)
Sarcasm	Ptácek et al. [165]	http://liks.fav.zcu.cz/sarcasm/	200,000	Yes (HAT)
	Ling and Klinger [112]	http://www.romanklinger.de/ironysarcasm	99,000	Yes (HAT)
	Amir et al. [5]	https://github.com/samiroid/CUE-CNN	11,541	Yes (HAT)
	Rajadesingan et al. [168]	Contact author	9,104	Yes (HAT)
	Ghosh and Veale [62]	https://bit.ly/31tMd8E	41,000	Yes (HAT)
	Bamman and Smith [8]	Contact author	19,534	Yes (HAT)
	Riloff et al. [181]	Contact author	175,000	Yes (MAT)
	Website (The sarcasm detector)	http://thesarcasmdetector.com	120,000	Yes (HAT)
	Github	https://github.com/topics/sarcasm-detection	-	Yes (HAT, MAT)
	Ghosh et al. [61]	http://alt.qcri.org/semeval2015/task11/	9,000	Yes (HAT)
	Oprea and Magdy [151]	https://github.com/silviu-oprea/isarcasm	4,484	Yes (MAT)
	Filatova [57]	Contact author	1,254	Yes (Manual)
Irony	Tsur et al. [208]	Contact author	66,000	Yes (Manual)
	Karoui et al. [96]	https://bit.ly/2Mxz6z8	38,262	Yes (HAT)
	Hee et al. [74]	https://github.com/Cyvhee/SemEval2018-Task3	3,834	Yes (MAT)
	Github	https://github.com/topics/irony-detection	-	Yes (HAT, MAT)
	Ghosh et al. [61]	http://alt.qcri.org/semeval2015/task11/	9,000	Yes (HAT)
	Filatova [57]	Contact author	1,254	Yes (Manual)
Satire	Burfoot and Baldwin [29]	Contact author	4,233	Yes (LTD)
	Rubin et al. [183]	Contact author	-	Yes (LTD)
Humor	Mihalcea and Strapparava [130]	Contact author	32,000	Yes (STD)
	Yang et al. [224]	Contact author	4,626	Yes (STD)
	Chen and Soo [35]	https://bit.ly/2N5YeMy	231,657	Yes (STD)
	Raz [174]	http://funtweets.com/ (Contact author)	-	Yes (STD)
Metaphor	Niculae and Danescu-Niculescu-Mizil [148]	http://vene.ro/figurative-comparisons/	1,400	Yes (LTD)
	Ghosh et al. [61]	http://alt.qcri.org/semeval2015/task11/	9,000	Yes (HAT)
Hyperbole	Troiano et al. [207]	Contact author	2,117	Yes (LTD)

the tweets. In Twitter, sarcasm-related hashtags (e.g., #sarcasm, #sarcastic, and #sarcasme) or non-sarcasm-related hashtags (e.g., #not, #politics, #education, and #humor) are used to create labeled datasets, assuming that the users are the best judge to mark their own tweets as sarcastic or non-sarcastic. Such hashtag-based datasets are reported in References [1, 8, 15, 20, 23, 56, 61, 87, 99, 111, 145, 168, 217]. Although the hashtag-based approach facilitates the creation of large-scale labeled datasets, there is always a question about the correctness of the hashtags mentioned by the users. It may mislead the whole training process due the usage of irrelevant sarcasm-related hashtags [85]. Fersini et al. [56] considered a hybrid approach in which hashtag-based labeled datasets are manually examined for the generation of more fine-grained and authentic datasets.

— *Tweets Metadata (TM)*: In addition to the approaches mentioned above, some researchers have considered tweet metadata for sarcasm dataset creation. Twitter metadata consists of information about a user, such as past tweets, location, re-tweets count, Twitter user

ID, and Twitter user name. For example, Rajadesingan et al. [168] considered 80 tweets of each user, apart from the labeled datasets collected using hashtags. These past tweets can be used for constructing features based on contrasting context and past sarcastic remarks. Khattri et al. [99] considered a tweet as sarcastic if contrasting sentiment words are present in it or it contrasts with the user's historical tweets in terms of sentiment. Named entity phrases from tweets within the users' timeline are searched to obtain true sentiment, and then historical sentiments are used to predict whether the user is sarcastic in the current tweet.

- *Long-text Datasets (LTD)*: Apart from Twitter, other online data sources are also considered in sarcasm-related studies. For example, Amazon product reviews dataset is used in References [42, 208], movie review datasets in References [133, 134], Internet Argument Corpus (IAC)[34] dataset in Reference [119], and discussion forums in References [87, 119]. Further, Reference [191] considered datasets from Instagram and Tumblr. The dataset used in Reference [208] can be obtained on request.
- *Others*: Apart from Twitter and LTD, Joshi et al. [88] used the "Friends"[35] dataset from the TV series, generated through a manual annotation process, where they consider utterances as sarcastic or non-sarcastic. Tepperman et al. [204] used call-center transcripts, and classified "yeah right" as a discriminator for sarcasm and non-sarcasm. Sarcastic and non-sarcastic excerpts were read and annotated by students in Reference [103]. These excerpts consist of longer narratives from booklets. As a guideline, three questions are given to the student annotators: Q1: "How likely is it that the speaker was being sarcastic?" Q2: "Why do you think so?" Q3: "How certain are you that the speaker was being sarcastic?." Q1 and Q3 were on a seven-point scale, whereas Q2 was free-form. Wallace et al. [215] created labeled sarcasm datasets from Reddit.[36] Mishra et al. [134] considered a manually annotated dataset from the Sarcasm society website.[37] Mishra et al. [133] also proposed a new kind of annotation technique where they recorded the eye movements of the manual annotators while reading the hashtag-based labeled tweets. These eye-tracking annotations provide supplementary annotation, and on the basis of that the authors proposed a predictive framework for sarcasm detection.

5.2 Irony-related Datasets

For irony, we have broadly classified the datasets used in the existing studies into two categories: (i) Twitter datasets and (ii) LTD.

- *Twitter datasets*: Since most of the researchers have considered irony detection as a binary classification problem for which annotated datasets are required, there have been different approaches to annotate Twitter datasets for irony detection.
  - —*MAT*: In a manual annotation approach, a tweet is manually labeled as ironic or non-ironic based on human judgment. For example, the datasets used in References [34, 51, 74, 213] are manually annotated.
  - —*HAT*: In Twitter, irony-related hashtags (e.g., #irony and #ironie) or non-irony–related hashtags (e.g., #not, #wtf, and #clinton) are used to create labeled datasets, assuming that the users are the best judge to mark their own tweets as ironic or non-ironic. Such hashtag-based irony datasets are reported in References [12–14, 16, 72, 96, 97, 179, 223].

[34] https://nlds.soe.ucsc.edu/iac.
[35] https://www.imdb.com/title/tt0108778/.
[36] https://www.reddit.com/.
[37] http://sarcasmsociety.com/.

- *LTD*: LTD for irony includes Amazon reviews, movie reviews, e-book articles, and newspaper articles. Carvalho et al. [31] used Portuguese newspaper data, and Reyes and Rosso [177] considered TripAdvisor[38] and Slashdot[39] data, in addition to Amazon customer reviews. Similarly, Tang and Chen [200] considered Plurk and Yahoo! data, and Wallace et al. [215] considered the popular social news website Reddit for irony detection.

5.3 Satire-related Datasets

Satire datasets used in the existing studies fall into two categories: (i) Twitter datasets and (ii) LTD.

- *Twitter Datasets*: For satire, there have been different approaches to annotate Twitter datasets, such as HAT, MAT, and TM.
  - *MAT*: In a manual annotation approach, a tweet is manually labeled as satire or non-satire, based on human judgment. For example, the datasets used in Reference [175] are manually annotated.
  - *HAT*: In Twitter, satire-related hashtags (e.g., #satire) or non-satire–related hashtags (e.g., #health, #food, and #news) are used to create labeled datasets, assuming that the users are the best judge to mark their own tweets as satire or non-satire. Such hashtag-based satire datasets are reported in Reference [175].
  - *TM*: Satirical and non-satirical Twitter accounts are used in References [9, 10, 186, 206] to create datasets for satire and non-satire categories.
- *LTD*: In the LTD category, online news articles [29], satire news articles [171], and Amazon product reviews [175] have been used by some researchers.

5.4 Humor-related Datasets

Humor recognition is typically considered as a binary classification task, i.e., a piece of text is either classified as humorous or non-humorous. Datasets used in humor-related studies can be broadly classified as (i) Twitter datasets, (ii) short-text datasets, and (iii) LTD.

- *Twitter Datasets*: Like aforementioned figurative language categories, humorous tweets are either based on HAT or MAT. The humorous tweets are mainly taken from comedian accounts (profiles), hashtags (i.e., #humor), and humor tweets repository available online[40] [174, 227].
- *Short-Text Datasets (STD)*: These include the 16,000 one-liners dataset [130], pun-of-the-day dataset [224], and the 231,657 short jokes dataset [35].
- *LTD*: The LTD include the "British National Corpus (BNC),"[41] "proverbs," and "Reuter's titles" in Reference [225], "Yelp reviews" in References [138, 150], and "news headlines" and "Wikipedia sentences" [18].

5.5 Simile-related Datasets

Simile datasets used in the existing studies fall into two categories: (i) Twitter datasets and (ii) LTD.

- *Twitter datasets:* Qadir et al. [166, 167] fetched tweets based on the "like," "as," and "than" keywords and annotated them manually to create MAT.

[38] https://www.tripadvisor.in/.
[39] https://slashdot.org/.
[40] http://www.funtweets.com/.
[41] http://www.natcorp.ox.ac.uk/.

- *LTD*: Amazon product reviews are considered in Reference [148]. Hao and Veale [69] prepared a dataset for ironic similes using Google API with different patterns, such as "as * as *" and "about as * as *" to extract snippets such as "*as hot as an oven*" and '*as strong as an ox*" and collected around 20K distinct similes.

5.6 Metaphor-related Datasets

Metaphor datasets include (i) Twitter datasets and (ii) LTD.

- *Twitter dataset*: Like sarcasm and irony, Twitter datasets for metaphor-related studies have been generated under the HAT category [61, 71, 95].
- *LTD*: It includes the British National Corpus [194] and the VU Amsterdam metaphor corpus [46, 47].

5.7 Hyperbole-related Datasets

As of now, there is only one hyperbole-related dataset, Troiano et al. [207]. They considered English exaggerations, literal paraphrases of the exaggerations as hyperbolic instances, and non-exaggerated sentences as non-hyperbolic instances.

6 EVALUATION METRICS AND VALIDATION APPROACHES

This section presents different performance evaluation metrics and validation approaches used to test the efficacy of figurative language detection methods.

6.1 Evaluation Metrics

Figurative language detection is generally considered as a classification problem, in which a given tweet is classified as either sarcasm or non-sarcasm, irony or non-irony, satire or non-satire, humor or non-humor, simile or non-simile, and metaphor or non-metaphor. As a result, metrics such as precision, recall, F-score, and accuracy are generally used for evaluation. These metrics have been used in References [5, 15, 21, 23, 42, 67, 69, 81, 82, 87, 88, 97, 99, 114, 119, 120, 137, 156, 163, 165, 166, 168, 181, 191, 208, 226].

These are defined using the concept of *True Positives* (TP) (i.e., the number of figurative language utterances identified as figurative language), *False Positives* (FP) (i.e., the number of normal utterances identified as figurative language), *False Negatives* (FN) (i.e., the number of figurative language utterances identified as normal), and *True Negatives* (TN) (i.e., the number of normal utterances identified as normal). *Precision* measures the correctness, whereas *recall* measures the completeness of any classification or information retrieval system. The harmonic mean of precision and recall is called the *F-score*, which is high when both precision and recall values are high. *Accuracy* measures the fraction of correct predictions. The *Precision*, *Recall*, *F-score*, and *Accuracy* are formally defined in Equations (1), (2), (3), and (4), respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Figurative language datasets are generally found to be skewed in nature. To deal with such datasets, *Area Under the Curve* (AUC) performs better than F-score and has been used in References [1, 106, 111, 121]. AUC estimates a combined measure of performance within the thresholds fixed for all possible classifications. However, some of the studies, such as Xu et al. [223] and Ghosh et al. [61], present evaluation results in the form of Mean Squared Error (MSE) and Cosine Similarity (CS). MSE measures predictive system performance and is generally used in optimization. Mathematically, it is defined in Equation (6), where Y and $\hat{Y}$ represent the actual and predicted values, respectively. CS is used to measure similarity between two documents represented as real-valued vectors A and B. It is formally defined in Equation (5).

$$Cosine\ (A, B) = \frac{A \cdot B}{|A| \cdot |B|} \tag{5}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y - \hat{Y})^2 \tag{6}$$

In addition, Qadir et al. [167] considered a statistical measure, Mean Reciprocal Rank (MRR), which is used to rank candidate results for an information retrieval query set (Q), as defined in Equation (7).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{7}$$

6.2 Validation Approaches

To validate the evaluation results, researchers have mainly considered cross-validation techniques (e.g., k-folds cross validation), which have been used in References [1, 15, 24, 35, 42, 67, 88, 116, 133, 134, 138, 143, 145, 165, 168, 181, 187, 193, 208, 210, 227]. In this approach, the dataset is partitioned into k parts, considering (k-1) parts for training and one part for testing. The process is repeated k times to ensure the testing of the model on each and every example within the dataset. Some authors have considered bootstrap sampling (also called 0.632 bootstrap) in which the training sets are drawn at random with replacement from the original data (containing on average 63.2% instances), and the remaining points comprise the testing sets (containing, on average, 36.8% instances).

Another approach is to use totally unseen test datasets. In this approach, categories of figurative language detection models are trained over a given training dataset and tested over new instances for which class labels are not known. Test dataset validation approach has been used in References [14, 23, 24, 49, 62, 69, 106, 111, 119–121, 139, 148, 152, 156, 179, 199, 201, 222].

7 FIGURATIVE LANGUAGE DETECTION APPROACHES

In this section, we present a review of existing literature on figurative language detection techniques. We present the different approaches for each FL category in line with the description of the datasets in Section 5 for the sake of better understanding.

7.1 Sarcasm Detection Approaches

The computational detection of sarcasm employs various Machine Learning (ML) techniques, which mainly include supervised learning techniques, such as Support Vector Machine (SVM), Naive Bayes (NB), Bayesian Networks (BNs), Maximum Entropy (ME), Random Forests (RF), Neural Networks (NNs), Logistic Regression (logR), K-Nearest Neighbor (K-NN), and Decision Trees (DT) [15, 23, 67, 88, 120, 143, 145, 165, 168, 187]. In addition, semi-supervised learning [42, 119,

208], rule-based techniques [125, 181], linguistic-based classification approaches [8, 111], ensemble learning [56, 117], and deep learning have also been considered for sarcasm detection [5, 45, 62, 161, 191, 226]. For the latter, the typical models include Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), Bi-Long Short-Term Memory networks (Bi-LSTMs) and Gated Recurrent Units (GRUs).

Table 3 presents a summary of the existing literature on sarcasm detection.

7.1.1 Supervised Approaches. Supervised learning is based on labeled training data in which the class label of each instance is given as input as the supervisory signal.

HAT: González-Ibáñez et al. [145] considered lexical (unigrams and dictionary-based) and pragmatics-based (smiley, frowning faces, and ToUser) features for automatic identification of sarcastic messages in positive- and negative-sentiment-bearing tweets. They applied SVM and logistic regression (LogR) for the classification task and found that SVM performs better than LogR. They observed that sarcasm detection is a difficult task for both humans and machine learning techniques due to the absence of explicit context markers.

Barbieri et al. [15] considered DT as the classification technique and included seven sets of features that consist of frequency, written-spoken words, intensity, structure, sentiment, synonyms, and ambiguity. They targeted the inner structure of sarcastic tweets using lexical features and avoided pattern-based features. They considered separation of sarcasm from irony as a future task.

Rajadesingan et al. [168] proposed a behavior-based model for sarcasm classification. The authors considered text expression, emotion, contrast, familiarity, and complexity features. They applied supervised learning techniques such as SVM, LogR, and DT to evaluate the model. They concluded that historical information, such as the author's past data, may help in sarcasm detection. Similarly, Wang et al. [217] used SVM^{hmm} on a Twitter dataset to compare sarcastic utterances to those utterances that show positive and negative sentiments without any use of sarcasm. Authors admitted that contextual clues play an important role in sarcasm detection, and they model it as a sequential classification task over a tweet and its contextual information.

Nguyen and Jung [146] proposed a figurative language identification method based on two models. The first one is a content-based approach, while the second one follows an emotional pattern-based approach. They observed that figurative language detection using statistical-based models produce good results. Muresan et al. [143] considered the effects of lexical and pragmatics features and applied various ML classifiers such as SVM, NB, and LogR. They extended their previous work reported in Reference [145]. They found that automatic classification can be as good as human classification. However, they admitted that performance is still weak and needs improvement. Bouazizi and Ohtsuki [24] considered sentiment, pattern, punctuation, syntactic, and semantic-based features to detect sarcastic utterances. The authors applied SVM, RF, ME, and K-NN classifiers. They focused on POS tags to extract patterns to characterize sarcasm in tweets for enhancing the performance of opinion mining and sentiment analysis–based systems.

Abulaish and Kamal [2] noticed that sarcasm can also be categorized into seven categories,[42] such as "*self-deprecating*," "*brooding*," "*deadpan*," "*polite*," "*obnoxious*," "*manic*," and "*raging*." They proposed a *self-deprecating sarcasm* detection approach using a two-layer approach. The first layer is used for filtration of 107,536 candidate self-around tweets from 151,283 preprocessed tweets. The second layer is composed of 11 features, i.e., 6 self-deprecating and 5 hyperbolic features. The task of the second layer is to classify a self-around tweet as self-deprecating sarcasm or non-self-deprecating sarcasm. They applied machine learning classifiers, such as NB, DT, and bagging, and observed that *self-deprecating sarcasm* is very commonly used in Twitter and deserves greater attention.

[42] http://edtimes.in/seven-types-sarcasm/.

Table 3. A Summary of the Existing Literature on Sarcasm Detection

Approach	Dataset	Literature	Feature	Dataset size	Eval. res.	Val. appr.	Classfication
Supervised	HAT	González-Ibáñez et al. [145]	pragmatics, unigrams, lexical	2,700	Accuracy: 0.75	5-fold C.V	Ternary
		Muresan et al. [143]	lexical, pragmatic	2,700	Accuracy: 0.78	5-fold C.V	Ternary
		Ptácek et al. [165]	**n-gram, word shape pattern, POS**	**200,000**	**F-scores: 0.94**	**5-fold C.V**	**Binary**
		Barbieri et al. [15]	written-spoken, frequency, intensity, structure, sentiments, synonyms, ambiguity	50,000	F-score : 0.62	10-fold C.V	Binary
		Bouazizi and Ohtsuki [24]	**punctuation, sentiment, syntactic, pattern, semantic**	**9,256**	**Accuracy: 0.83**	**10-fold C.V, Test dataset**	**Binary**
		Abulaish and Kamal [2]	self-deprecating, hyperbolic	107,536	F-score: 0.94	10-fold C.V	Binary
	MAT	Lunando and Purwarianti [120]	unigrams, negativity, interjection words, question words	1,280	Accuracy: 0.54	Test dataset	Binary
		Tungthamthiti et al. [210]	sentiment score, punctuation, n-grams	50,000	Accuracy: 0.79	10-fold C.V	Binary
		Bouazizi and Ohtsuki [23]	sentiment, syntactic, punctuation, pattern	21,200	Accuracy: 0.83	Test dataset	Binary
		Gupta and Yang [67]	sociolinguistics, affect, cognitive	30,848	F-score: 0.60	10-fold C.V	Binary
		Samonte et al. [187]	**lexical, hyperbolic, pragmatics**	**12,000**	**Accuracy: 0.98**	**5-fold C.V**	**Binary**
	LTD	Justo et al. [92]	semantic, statistical, linguistic, lex-ical	9,226	Accuracy: 0.68	10-fold C.V	Binary
	Other	Joshi et al. [88]	conversation context, speaker co-ntext, lexical	17,338	F-score: 0.84	5-fold C.V	Binary
	MAT, HAT	Abercrombie and Hovy [1]	Twitter, authors, audience, historical, environment	4,480	AUC: 0.60	5-fold C.V	Binary
	HAT, TM	Rajadesingan et al. [168]	text expression, complexity, emo-tion, contrast, familiarity	9,104	Accuracy: 0.83	10-fold C.V	Binary
	HAT, LTD	Parde and Nielsen [156]	polarity, subjectivity, BOW	6,252	F-scores: 0.59 (Twitter), 0.78 (Amazon)	Test dataset	Binary
	Other, HAT, LTD	Mishra et al. [134]	implicit and explicit incongruity, cognitive (Gaze related), lexical	1,000	F-score: 0.75	10-fold C.V	Binary
		Mishra et al. [133]	**gaze, textual**	**1,000**	**F-score: 0.93**	**5-fold C.V**	**Binary**
Semi-Supervised	LTD, HAT	Davidov et al. [42]	syntactic, pattern	66K (Amazon), 5.9M (Twitter)	F-score: 0.83	5-fold C.V	Binary
	LTD	**Tsur et al. [208]**	**syntactic, pattern, punctuation**	**66,000**	**F-score: 0.78**	**5-fold C.V**	**Binary**
		Lukin and Walker [119]	sarcastic and nasty patterns	10,003	F-score: 0.69	Test dataset	Binary
Rule-Based	HAT	Bharti et al. [20]	interjections, negative sentiment and positive situation, hyperbole	56,500	F-score: 0.90	N.A	Binary
		Bharti et al. [21]	**interjections, parsing**	**1.45M**	**F-score: 0.97**	**N.A**	**Binary**
	MAT	Maynard and Greenwood [125]	Twitter-hashtag tokenizer	400	Precision: 0.91	N.A	Binary
	MAT, HAT	Riloff et al. [181]	positive verbs, negative phrases	178,000	F-score : 0.51	10-fold C.V	Binary
	HAT, TM	Khattri et al. [99]	past tweets, contrast related tweets	10,278	F-score: 0.88	N.A	Binary
Linguistics	HAT	Liebrecht et al. [111]	intensifier, exclamation, emotional marks, ngrams	3.67M	AUC : 0.79	Test dataset	Binary
		Bamman and Smith [8]	author's profile informa- tion, historical sentiment	19,534	Accuracy: 0.85	10-fold C.V	Binary
		Kunneman et al. [106]	**punctuations, emoticons, unigram, bigram, trigram**	**2.65M**	**AUC: 0.85**	**Test dataset**	**Binary**

(Continued)

Table 3. Continued

Approach	Dataset	Literature	Feature	Dataset size	Eval. res.	Val. appr.	Classfication
Deep-Learning	HAT	Ghosh and Veale [62]	BOW, POS	41,000	F-score: 0.92	Test dataset	Binary
		Amir et al. [5]	contextual, tweets-response, author, audience	11,541	Accuracy: 0.87	10-fold C.V	Binary
		Poria et al. [161]	**sentiment, emotion, personality**	**200,000**	**F-score: 0.97**	**Test dataset**	**Binary**
		Zhang et al. [226]	neural, contextual	9,104	Accuracy: 0.94	10-fold C.V	Binary
	HAT, Other	Schifanella et al. [191]	visual semantics, n-grams, subjectivity, textual	4,050 (Twitter), 20,000 (Instagram), 20,000 (Tumblr)	Accuracy: 0.89	N.A	Binary
Ensemble	HAT	Fersini et al. [56]	BOW, POS	8,000	F-score: 0.83	10-fold C.V	Binary
	HAT, LTD	Liu et al. [117]	syntactic, lexical, Rhetoric	69,426	AUC: 0.89	10-fold C.V	Binary
Fuzzy	HAT	Mukherjee and Bala [142]	Content, Function, POS	2,000	ACC: 0.65	10-fold C.V	Binary

The bold entries show the best performing results among the various state-of-the-arts using the same approach and dataset category.

MAT: Tungthamthiti et al. [210] considered n-gram, punctuation, special symbol, and sentiment score as features and applied an SVM classifier. They focused on mainly sentiment analysis, concept level and common-sense knowledge, coherence, and classification. Similar to Reference [145], they observe that sarcasm is tough to diagnose, and it depends mainly on the common sense knowledge and existing context in an instance/tweet. Further, Bouazizi and Ohtsuki [23] considered sentiment-, punctuation- syntactic-, and pattern-related features. They used NB, SVM, and ME classifiers. They showed the importance of detecting sarcasm in tweets to enhance sentiment analysis and opinion mining. Abercrombie and Hovy [1] emphasized context-based sarcasm detection. The authors judged the performance of humans and machines to recognize sarcasm. They concluded that class balance and dataset size should be taken into account when designing sarcasm detection systems.

TM: Mishra et al. [133, 134] consider cognitive (Gaze) features to detect eye movement of human readers. Apart from that, they also applied lexical features. Further, they used SVM, NB, NN, and multi-instance LogR (MILR) classifiers. Using cognitive features, such as eye movement, serves as supplementary annotation for sarcasm detection. Likewise, Mishra et al. [133] considered gaze behavior of readers to understand sarcasm. The authors extracted lexical gaze (skip count, regression count, fixation count) and textual (i.e., interjections, punctuation, positive words, negative words) features. They considered cognition-cognizant techniques involving eye-tracking as a promising approach for sarcasm detection and interpretation.

Joshi et al. [86] developed a browser-based system for sarcasm detection and generation. The sarcasm generation module provided by the authors is a chat-bot that replies in a sarcastic way to a user input.

LTD: Justo et al. [92] considered statistical, linguistic, lexical, and semantic features to detect nastiness and sarcasm from online communications using rule-based and NB classifiers. They observed that linguistic and semantic information are good indicators of sarcasm. Gupta and Yang [67] proposed affect-, cognition-, and sociolinguistics-related features and trained an SVM classifier to detect sarcastic tweets. They developed a two-level cascade classification system and observed that sarcasm detection derived features consistently benefited key sentiment analysis evaluation metrics. Das and Clark [39] proposed sarcasm detection on Facebook[43] data. Their work

[43] https://www.facebook.com/.

on sarcasm detection considers various types of content available in Facebook posts, such as text, images, and user interactions.

Other Data: Joshi et al. [88] considered lexical (unigram), conversation context (action words, sentiment score, previous utterance sentiment score value), and speaker context (speaker name and speaker-listener pair) features and applied sequence labeling techniques such as SVM^{hmm} [4] and SEARN [41] for sarcasm detection from the TV series *Friends*. They observed the efficacy of sequence labeling techniques for sarcasm detection in dialogues.

Joshi et al. [87] considered lexical, pragmatics, and explicit and implicit incongruity as features and applied SVM [33] on discussion forum posts and tweets. They reported how context incongruity theory is useful for sarcasm detection. Parde and Nielsen [156] analyzed domain-general sarcasm detection performance on Twitter and Amazon product reviews. They analyzed common types of behavior for sarcasm across domains. They considered polarity, subjectivity, and BOW features. They applied NB classifier. Agrawal and An [3] highlighted affective content and its effectiveness for word representations to detect sarcasm. They considered Twitter, reviews, and discussion forum posts, and observed that affective representation showed better results on short texts, such as Twitter.

Besides the research works discussed above, there are some articles that aim at detecting mixed categories of figurative language such as sarcasm, irony, satire, and metaphor [11, 53, 56, 71, 89, 95, 100, 112, 127, 146, 147, 171, 198, 216].

7.1.2 Semi-supervised Approaches. Semi-supervised learning lies between supervised learning (where class labels of instances are known) and unsupervised learning (where class labels of instances are not known). For training, unlabeled data are used along with a small amount of labeled data, and many researchers have explored this approach for sarcasm detection.

LTD: Tsur et al. [208] proposed the *semi-supervised sarcasm identification* (SASI) algorithm to identify sarcasm in Amazon product reviews. The algorithm consists of two modules: (i) semi-supervised pattern acquisition and (ii) sarcasm classification. First, the authors manually annotate and label a small set of sentences with a score from 1 to 5, where a 5 indicates a fully sarcastic sentence and a 1 indicates complete absence of sarcasm. Thereafter, they construct feature vectors for each labeled sentence in the dataset and build a classification model for assigning scores to unlabeled sentences. The authors generated reviews from Amazon for training and used syntactic, pattern, and punctuation features to learn a K-NN classifier.

High-frequency words are those words whose corpus frequency is more than F_H. Content words are those words whose corpus frequency is less than F_C, where F_H and F_C are used as threshold values.

Davidov et al. [42] adopted the same semi-supervised approach and used SASI for sarcasm detection in a Twitter dataset containing around 6M tweets and the same Amazon product reviews containing 66K documents. Based on syntactic and pattern-based features, the authors achieved good results for both datasets. Lukin and Walker [119] identified sarcastic and nasty patterns using bootstrapping on the Internet Argument Corpus (IAC), which includes categories such as sarcastic versus non-sarcastic, nasty versus nice, and rational versus emotional. They generated a seed set of nasty/sarcastic patterns using Amazon Mechanical Turk, derived from a labeled dataset. Thereafter, a bootstrapping process is applied over the unlabeled dataset to learn new extraction patterns for sarcasm/nasty classification.

7.1.3 Rule-based Approaches. Rule-based approaches provide information using a set of rules, which are either constructed by domain experts or via automatic rule inference systems.

MAT: Riloff et al. [181] proposed a bootstrapped lexicon-based approach to recognize sarcasm from Twitter, targeting phrases based on positive verb sentiment and negative situation. Consider

the example, "Absolutely *adore* it when my *bus is late,*" taken from Reference [181]. Here, the sarcasm occurs due to contrast of positive word "adore" with a negative phrase "bus is late." Maynard and Greenwood [125] applied a rule-based approach to figure out sentiment in sarcastic sentences. The authors applied a Twitter hashtag tokenization technique to detect sentiment and sarcasm in hashtags. Consider the example, "I am not happy that I woke up at 5:15 this morning.. *#greatstart #sarcasm*," taken from Reference [125]. Here, sarcasm lies in the hashtag content #greatstart. The rest of the sentence, excluding hashtags, in this example is negative.

HAT: Bharti et al. [20] applied a rule-based approach to detect sarcasm in Twitter texts. The authors applied interjections, intensifier, hyperbole, and phrase (negative sentiment and positive situation) features. They proposed two algorithms: The first algorithm forms a parse tree for sentences and identifies phrases based on situations that indicate sentiments. If in a positive sentence there exists a negative phrase, then such sentences are recognized as sarcastic. The second algorithm considers tweets starting with an interjection as sarcastic. They considered three human annotators to validate the training set for both of the proposed algorithms. Bharti et al. [21] proposed a Hadoop-based framework for sarcasm detection in real-time Twitter streaming data, applying interjections and parsing-based features.

TM: Khattri et al. [99] proposed a rule-based approach in which sentiment from the past tweets of a user is used for sarcasm detection. In addition, they proposed a contrast-based predictor in which sentiment contradictions in the target tweets are monitored. They conclude that text written by an author in the past to identify sarcasm in a piece of text opens a new direction of research. Parmar et al. [157] proposed a Hadoop-based framework for sarcasm detection and considered lexical and hyperbole features for sarcasm detection.

7.1.4 Linguistic-based Approaches. The scientific study of language is related to the term "linguistics." However, "computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena."[44] There are relatively few studies that follow a linguistics approach for sarcasm detection.

HAT: Bamman and Smith [8] considered tweet-related features (POS, pronunciation, intensifier), author features (profile information, historical topics, historical sentiments), audience features, and environment features and applied a binary logistic regression technique. They observe that the inclusion of #sarcasm is not exactly a direct pointer for sarcasm tweets. Kunneman et al. [106] considered unigram, bigram, trigram, punctuation, and emoticon features and applied the Winnow classification model [113]. They considered the role of intensifiers in sarcastic texts.

7.1.5 Deep Learning-based Approaches. Deep learning is a powerful machine learning technique that is particularly based on data representation learning. Recently, deep learning has emerged as a popular technique for natural language processing and artificial intelligence problems. In the past few years, the state-of-the-art accuracy results obtained using deep learning models have attracted many researchers. In traditional machine learning models, a great amount of time is taken for feature engineering process, whereas deep learning models do not require hand-crafted features. Instead, they automatically learn different representations from data itself. Deep learning models consist of multiple processing layers to learn data representations and produce excellent results. Apart from image data, they are highly effective for text data processing, including figurative language detection [109].

HAT: Ghosh and Veale [62] proposed a neural network–based semantic model composed of Deep NNs, CNNs, and LSTMs for sarcasm detection. They applied BOW and POS-based features.

[44] https://www.aclweb.org/archive/misc/what.html.

They obtained good results on training and test datasets as compared to the recursive SVM approach. They observed the usefulness of neural network–based semantic modeling for sarcasm detection. Amir et al. [5] proposed a content and user embedding–based CUE-CNN model to extract sarcastic utterances. Instead of using hand-crafted features, their model automatically learns embeddings for content and users and is used in concert with lexical signals for sarcasm detection.

Poria et al. [161] considered both balanced and unbalanced datasets from Ptácek et al. [165] for training and a dataset from the sarcasm detector website for testing. They developed a model using a pre-trained CNN for extracting sentiment, emotion, and personality features for sarcasm detection. Zhang et al. [226] considered syntactic and semantic features from Twitter and applied a bidirectional gated RNN. In addition, the authors applied a pooling neural network to obtain the contextual features from historical tweets. They used the Rajadesingan et al. [168] dataset for training. Schifanella et al. [191] implemented a novel multimodal system using both textual and visual data from three social media platforms (Twitter, Instagram, and Tumblr) for sarcasm detection. The authors considered visual semantics, subjectivity (i.e., number of first-person pronouns, third-person pronouns, and passive constructs), n-grams, and textual features, and applied CNN and SVM for sarcasm detection.

Das and Clark [40] applied sarcasm detection on Flickr images using a CNN. Dubey et al. [45] proposed the task of converting sarcastic into non-sarcastic interpretation. They used a rule-based, statistical machine translation and deep learning-based approach employing an encoder-decoder, pointer generator, and attention network. They mainly used negation to get non-sarcastic interpretation of the sarcastic texts.

7.1.6 Ensemble Learning Approaches. Ensemble learning is based on multiple learners, and these are trained in such a way that they solve a problem together. Fersini et al. [56] consider pragmatics (emotions, onomatopoeic, punctuation) and POS-tags features and apply a Bayesian model-averaging approach, which outperforms the majority voting mechanism and other ensemble learning methods.

7.1.7 Fuzzy Clustering–based Approaches. Fuzzy clustering (soft clustering) allows each data point to lie in multiple clusters. Mukherjee and Bala [142] proposed a fuzzy clustering approach using applied content words, function words, POS tags, and POS n-grams features. Content words are those words that have meaning outside the sentence, such as "dog," and "college." Function words are those words that have no sense or meaning outside the sentence boundary, such as "and," "the," and "not." They used the Fuzzy C-Means (FCM) algorithm to detect sarcasm, although the results obtained using FCM did not show better results in comparison to NB classification due to the small dataset of only 2K tweets.

7.1.8 Multilingual Studies for Sarcasm Detection. Apart from English, some researchers have considered other languages for sarcasm detection. Ptácek et al. [165] considered *English* and *Czech* tweets for sarcasm detection using n-grams and POS-based features and applied SVM and ME classifiers. They concluded that in-depth linguistic insights would be helpful for better understanding of sarcasm on social media. Liu et al. [117] proposed an ensemble learning approach to deal with the class imbalance problem in *Chinese* datasets. Liebrecht et al. [111] considered intensifiers, n-grams, exclamations, and emotional marks features and applied the balanced Winnow [113] linguistic classification technique for multi-label classification in a *Dutch* language dataset. They observed that different markers—such as hashtags used across different languages—are often used to mark sarcasm instances. Lunando and Purwarianti [120] detected sarcastic utterances in *Indonesian* tweets. They considered unigrams, negativity, number of interjection words, and question words, and applied SVM, NB, and ME for sarcasm detection. They observed that the negativity

feature indicates sentiment value, whereas the interjection feature represents lexical aspects. Bharti et al. [19] considered sarcasm detection in *Hindi* tweets related to news context. They compared a set of keywords for both input tweet and related news. They observed that news articles contain neutral sentiment, and if the orientation of news and tweet are not the same in terms of polarity, then the user is trying to negate this temporal fact, and the given input tweet contains sarcasm. Samonte et al. [187] proposed sentence-level sarcasm detection in datasets containing tweets in Austronesian (a language spoken in the Philippines) and English. They considered lexical, pragmatics, and hyperbole features and applied SVM, NB, and ME for sarcasm detection. They concluded that annotated and balanced datasets are important for sarcasm classification.

7.2 Irony Detection Approaches

Irony detection approaches mainly employ supervised and deep learning–based techniques that are discussed in the following subsections and are summarized in Table 4.

7.2.1 Supervised Approaches. Mostly, HAT, and LTD are used in supervised approaches for irony detection tasks.

HAT: Reyes et al. [179] considered features such as ambiguity, polarity, emotional scenarios, and unexpectedness and applied DT for classification. Their model is based on textual features covering two dimensions—representativeness and relevance. Their results provide valuable insight regarding the creative and positive usage of two figurative language categories—irony and humor. Later on, Reyes et al. [180] applied the same approach as Reference [179], but considered unexpectedness, emotional scenarios, style, and signature features for DT and NB classifiers. However, combining all these features performed better.

Barbieri and Saggion [12] proposed four distinct topics—*education*, *humor*, *politics*, and *irony*—and considered frequency, written-spoken, intensity, structure, sentiment, synonyms, and ambiguity feature groups for RF and DT classifiers. The authors found that ambiguity is the least discriminative and proposed considering more discriminating features for irony detection using supervised approaches. Similarly, References [13, 14] considered the same datasets and feature sets. De Freitas et al. [43] considered features such as emoticons, laughter expressions, adjectives, quotation marks, and demonstrative pronouns and applied a linguistic approach for irony detection. Farías et al. [51] developed the *emotIDM* model for irony detection. They considered structural, affective, and emotional features and applied NB, DT, and SVM classifiers on the datasets from References [12, 15, 180]. Based on information gain, they concluded that affective features are more discriminating to distinguish ironic and non-ironic tweets.

LTD: Reyes and Rosso [177] considered Amazon reviews for irony detection. They identified various n-grams, POS n-grams, and profiling (funny, positive/negative, affective, and pleasantness) features and learned SVM, DT, and NB classifiers for irony detection. They considered two goals in their evaluation—feature relevance and capability of finding ironic documents. In addition, the authors planned to manually annotate irony instances in the future. Reyes and Rosso [176] considered a set of customer reviews, which are found as ironic. These reviews triggered a chain reaction once they became viral, both on social and mass media. Similar to Reference [177], they used six features to design a model for characterizing irony.

Reyes and Rosso [178] collected data on movie reviews, book reviews, and news articles from Burfoot and Baldwin [29]. They considered textual features, such as pointedness, imagery, activation, temporal imbalance, temporal compression, pleasantness, counterfactuality, and contextual imbalance for irony detection. They admit that combining all these features provides a valuable linguistic inventory for irony detection task. They reported two kinds of results—isolated sentences and entire documents, based on the annotations using two key strata. The first strata considered

Table 4. A Summary of the Existing Literature on Irony Detection

Approach	Dataset	Literature	Feature	Dataset size	Eval. res.	Val. appr.	Classfication
Supervised	LTD	Carvalho et al. [31]	demonstrative, determiners, onomatopoeic expressions, punctuation quotation marks, diminuitive forms, interjections	258,211	Precision: 0.45-0.85	N.A	Multi-class
		Reyes and Rosso [177]	**POS n-grams, profiling, POS**	**11,861**	**F-score: 0.89**	**10-fold C.V**	**Binary**
		Reyes and Rosso [176]	POS n-grams, profiling, POS	8,861	F-score: 0.78	10-fold C.V	Binary
	HAT	**Reyes et al. [179]**	**unexpectedness, polarity, emotional sce-nario, morphosyntactic ambiguity, str-uctural ambiguity, semantic ambiguity**	**50,000**	**F-score: 0.93**	**Test dataset**	**Binary**
		Reyes et al. [180]	unexpectedness, emotional scenarios, style, signatures	40,000	F-score: 0.76	10-fold C.V	Binary
		Barbieri and Saggion [12]	written-spoken, frequency, intensity, struct-ure, sentiments, synonyms, ambiguity	40,000	F-score: 0.88	10-fold C.V	Binary
		Barbieri and Saggion [13]	written-spoken, frequency, intensity, struct-ure, sentiments, synonyms, ambiguity	40,000	F-score: 0.75	10-fold C.V	Binary
		Barbieri and Saggion [14]	written-spoken, frequency, intensity, struct-ure, sentiments, synonyms, ambiguity	40,000	F-score: 0.75	Test dataset	Binary
		Karoui et al. [97]	surface, opposition, sentiment, shifter, sen-timent shifter	6,742	F-score: 0.86	10-fold C.V	Binary
		Taslioglu and Karagoz [201]	smiley, questions and exclamation marks, full stop, frowns faces, sentiment scores gaps	600	F-score: 0.73	Test dataset	Binary
	MAT	Charalampakis et al. [34]	spoken, lexical, emoticons, rarity	44,438	Precision: 0.83	10-fold C.V	Binary
	HAT, MAT	Farías et al. [51]	structural, affective, emotional	214,978	F-score: 0.96	N.A	Binary
Deep-Learning	LTD	Ravi and Ravi [172]	syntactic, semantic, and psycho-linguistic	1,022,171	AUC: 0.99	N.A	Binary
	HAT	Huang et al. [79]	linguistics	N.A	N.A	N.A	Binary
	HAT, MAT	Hee et al. [74]	handcrafted, word embedding	4,618	F-score: 0.71	Test dataset	Binary, Multi-class
	HAT, LTD	Zhang et al. [228]	sentiment features	121,026	F-score: 0.99	Test dataset	Binary

The bold entries show the best performing results among the various state-of-the-arts using the same approach and dataset category.

the whole sentence to be ironic or not on the basis of its content, whereas the second strata considered the context in each sentence to determine whether the document containing it would be regarded as being ironic or not.

7.2.2 Deep Learning–based Approaches. Recently, some deep learning–based approaches have been considered for irony detection.

HAT: Huang et al. [79] considered deep learning models such as RNN, CNN, and attentive RNN for irony detection. They highlight the importance of attention mechanism as an important linguistic clue for detecting ironic instances.

LTD: Ravi and Ravi [172] considered irony detection using syntactic, semantic, and psycho-linguistic features and applied Doc2Vec[45] word embedding. The authors observe that pre-trained word embeddings, such as Doc2Vec, and psycho-linguistic features are very helpful for irony classification. Zhang et al. [228] considered incongruity, which plays an important role in irony detection. They applied transfer learning–based approaches and used sentiment knowledge to improve the attention mechanism of RNNs for capturing hidden incongruity patterns. They reported two findings—first, sentiment knowledge from external resources is good for irony, and second, transferring deep sentiment features are effective to obtain implicit incongruity.

7.2.3 Multilingual Studies for Irony Detection. For irony detection, researchers have also considered languages other than English. Carvalho et al. [31] considered Portuguese language text data for irony detection. They considered Portuguese newspaper content and employed punctuations, interjections, diminutive forms, verb morphology, cross-constructions, quotation marks, and onomatopoeic expressions as features and applied a dictionary lookup for named entity recognition using a named entity lexicon. Diminutive forms are used to express positive sentiments and verb morphology is used to indicate pronouns as a way of expression in ironic texts in Portuguese. In cross-constructions, adjectives relate to the noun that is modified using prepositions. Onomatopoeic expressions are related to internet slang, such as "ah," "eh," and "hi." Quotation marks features are used in ironic content to put emphasis in text.

Bosco et al. [22] considered two important features, *polarity reversing* and *emotion expression*, for irony detection in two Italian language corpora; namely, *TWNews* and *TWSpino,* containing political tweets. Similarly, Basile et al. [16] considered Italian tweets to detect irony detection. They used word-based, syntactic, and semantic features. Further, Karoui et al. [97] considered pragmatics context as an indicator for irony detection and identified features such as surface, shifter, semantic, sentiment shifter, and opposition in French, using an SVM classifier. Stranisci et al. [197] presented an annotated Italian linguistic resource for sentiment analysis and irony. Tang and Chen [200] constructed an irony corpus in Chinese and extracted patterns for irony using a bootstrapping approach. They consider patterns, such as "degree adverbs followed positive adjective," "positive adjective with high intensity words," "positive noun with high intensity," "the use of very good," and "presence of negative adjective," as indicators for irony.

Charalampakis et al. [34] considered Greek political tweets and presented a comparison of supervised and semi-supervised techniques. They considered spoken, lexical, emoticons, and rarity features for classification. Hee et al. [72] retrieve English and Dutch language tweets using the #irony hashtag. They found that *contrasting evaluation* is a key indicator for irony detection. Contrasting evaluation can be present in an instance in the form of opposition, hyperbole, or an #irony hashtag. Karoui et al. [96] focused on pragmatic behavior to detect irony in English, Italian, and French language tweets. Taslioglu and Karagoz [201] considered Turkish and English tweets and considered features such as exclamation-, question-, and quotation-marks, sentiment gap scores, smileys, frowns (a.k.a. negative smileys), and diminutive forms. Ortega-Bueno et al. [153] considered irony detection in tweets and news comments over three Spanish variants.

Besides the research works discussed above, other approaches aim at detecting mixed categories of figurative language such as irony, sarcasm, satire, simile, and metaphor [11, 53, 56, 57, 61, 69, 71, 89, 95, 100, 112, 127, 146, 147, 171, 198, 205, 216].

Some works do not provide any evaluation metrics, but they can still provide directions for future research. For example, the approach of Reference [22] can be used with deep learning models to deal with *polarity reversing* and *emotion expression* features. Similarly, Hee et al. [72] discuss

[45] https://radimrehurek.com/gensim/models/doc2vec.html.

Table 5. A Summary of the Existing Literature on Satire Detection

Approach	Dataset	Literature	Feature	Dataset size	Eval. res.	Val. appr.	Classfication
Supervised	LTD	Burfoot and Baldwin [29]	headlines, profanity, BNS, slang	4,233	F-score: 0.79	Test dataset	Binary
		Rubin et al. [183]	**predictive**	**360**	**F-score: 0.87**	**10-fold C.V**	**Binary**
		Stöckl [196]	Not available	60,000	F-score: 0.76	Test dataset	Binary
	TM	Barbieri et al. [10]	POS, frequency, synonyms, characters, sentiments, ambiguity	6,533	F-score: 0.85	Test dataset	Binary
		Barbieri et al. [9]	frequency, synonyms, charac-ture, sentiments, ambiguity	34,281	F-score: 0.80	5-fold C.V, Test dataset	Binary
		Salas-Zárate et al. [186]	**psycholinguistic**	**20,000**	**F-score: 0.85**	**10-fold C.V**	**Binary**
	TM, LTD	Thu and Nwe [206]	word-based, emotion, sentiment	57,702	F-score: 0.80	10-fold C.V	Binary
	HAT, MAT, LTD	Reganti et al. [175]	n-grams, sensicon, lexical, sentiments amplifiers	13,254	F-score: 0.79	5/10-fold C.V	Binary
Deep-Learning	LTD	**Sarkar et al. [189]**	**syntax**	**186,549**	**F-score: 0.91**	**Test dataset**	**Binary**
		Dutta and Chakraborty [48]	sequence label sentiment	N.A	N.A	N.A	Binary

The bold entries show the best performing results among the various state-of-the-arts using the same approach and dataset category.

contrasting evaluation, which can be used for irony detection based on the concept of polarity contrast.

7.3 Satire Detection Approaches

As described in Reference [10], "*satire is a form of communication where humor and irony are used to criticize someone's behavior and ridicule it.*" Hence, satire is present in both sarcasm and irony. Like irony detection, satire detection approaches mainly employ supervised and deep learning–based techniques that are discussed in the following subsections. Table 5 presents a summary of the existing literature on satire detection.

7.3.1 Supervised Approaches. Supervised approaches are mainly applied over LTD for satire detection.

LTD: Burfoot and Baldwin [29] considered true news documents to identify satirical news articles. To generate a corpus using real and satirical news articles, they utilized bi-normal separation and lexical (headlines, profanity, slang) features and applied an SVM classifier. They applied two feature-weighting methods: (i) binary feature weighting and (ii) bi-normal separation feature scaling. In binary feature weighting, the same weight is assigned for all features regardless of whether they appear in an article once or more. However, the bi-normal separation feature scaling generates the highest weight for strongly correlated features that belong to either the negative or the positive class and lesser weight to features that occur evenly across the training instances. Equation (8) presents the formula to determine the weight of a feature f using the bi-normal separation feature scaling, where F^{-1}, TPR, and FPR indicate the inverse normal cumulative distribution function, true positive rate, and false positive rate, respectively.

$$weight(f) = |F^{-1}(TPR) - F^{-1}(FPR)| \tag{8}$$

Rubin et al. [183] considered five predictive features: absurdity, humor, grammar, negative affect, and punctuation, and applied an SVM-based classification approach. The authors collected 360 news articles from Canadian and US newspapers as corpus. After combining three out of five features (absurdity, grammar, and punctuation), their system yields good results. They observed that the BNS feature scaling is good for satire detection, as it retains a high precision.

Reganti et al. [175] utilized datasets from tweets, Amazon product reviews [57], and newswire articles (English Gigaword Corpus) [29]. They considered baseline (n-grams), lexical, sentiment amplifier, and speech act group of features. They observed that the usage of an ensemble classifier produces good results. Thu and Nwe [206] considered emotional features to classify satire and non-satire from news articles, Amazon product reviews, and news tweets, also observing good results for Twitter data through an ensemble classifier. Thu and Nwe [205] employed a satire detection model using emotion-related features, which they found useful for satire detection. Stöckl [196] considered satire detection using linear SVM and logR over the datasets containing news articles and satire website news. They also discussed satire detection along with other figurative language categories, such as *sarcasm*, *irony*, and *humor*. They noticed that non-linear kernels in SVM give poor results due to over-fitting.

7.3.2 Deep Learning–based Approaches. Recently, some deep learning–based approaches have been reported for irony detection in LTD.

LTD: Sarkar et al. [189] proposed deep learning–based techniques such as CNN, LSTM, and GRU to detect satire at both sentence and document levels. They concluded that fine-grained sentence-level analysis provides an in-depth insight into the phenomenon of satire; in particular, the presence of few key sentences, including the last sentence, is important for satire detection. Dutta and Chakraborty [48] determined an article as satire by using linguistic and machine learning tools. They extracted opinion expressions from token-level sequence-labeling of sentiments using a deep RNN from different-length text corpora.

7.3.3 Multilingual Studies for Satire Detection. Apart from English, some researchers have also considered other languages for satire detection. Barbieri et al. [10] considered advertisement of satirical news from tweets in Spanish using a satirical model based on Barbieri and Saggion [12]. They considered frequency, ambiguity, POS, synonyms, sentiments, characters, and slang words as features and employed two balanced binary classification experiments. They reported that cross-user account experiments provide good results. In such experiments, tweets in training and test datasets are not generated by the same Twitter accounts.

Barbieri et al. [9] introduced an automatic satirical news detection technique from tweets in English, Spanish, and Italian. They considered word-based (lemma, bigrams, skip-1,2,3 grams), frequency (rarest word frequency, frequency mean, frequency gap), synonyms, ambiguity, POS, sentiments, and punctuation features. The word-based features are used to capture common word-patterns. A binary classification approach was employed to classify satirical and non-satirical tweets, and they tested the performance of the system on both monolingual and cross-language experiments. Salas-Zárate et al. [186] proposed a psycho-linguistics approach. The authors collected a corpus of satirical and non-satirical news from Twitter's Mexican and Spanish accounts. They considered a wide variety of psychological and linguistic features and extracted those using LIWC.

7.4 Humor Recognition Approaches

Humor recognition approaches mainly employ supervised and deep learning–based techniques that are discussed in the following subsections and summarized in Table 6.

Table 6. A Summary of the Existing Literature on Humor Recognition

Approach	Dataset	Literature	Feature	Dataset size	Eval. res.	Val. appr.	Classfication
Supervised	STD, LTD	**Mihalcea and Strapparava [130]**	**stylistic, content**	**32,000**	**Accuracy: 0.96**	**10-fold C.V**	**Binary**
		Mihalcea and Pulman [128]	human-centeredness, neg-ative polarity	34,250	Accuracy: 0.96	10-fold C.V	Binary
	MAT, STD	Zhang and Liu [227]	phonetic, pragmatic, aff-ective	3,000	Accuracy: 0.847, F-score: 85	10-fold C.V	Binary
	LTD	Morales and Zhai [138]	content, ambiguity, alli-teration	1.6M	Accuracy: 0.85	5-fold C.V	Binary
	STD	Yang et al. [224]	phonetic, incongruity, ambiguity, interpersonal effect	36,828	F-score: 0.85	10-fold C.V	Binary
		Zhang et al. [225]	contextual, subjectivity, affective polarity	16,000	F-score: 0.85	N.A	Binary
		Liu et al. [115]	**phonetic, incongruity, syntactic**	**52,000**	**F-score: 0.92**	**10-fold C.V**	**Binary**
		Beukel and Aroyo [18]	homophones, ambiguity, homograph	44,652	Accuracy: 0.91	10-fold C.V	Binary
		Liu et al. [116]	sentiment conflict, sent-iment transition	20,000	F-score: 0.82	10-fold C.V	Binary
		Khandelwal et al. [98]	content, BOW, n-grams	3,453	Accuracy: 0.69	10-fold C.V	Binary
		Ermilov et al. [49]	**lexical, structural, BOW**	**47,000**	**Accuracy: 0.88**	**Test dataset**	**Binary**
Deep-Learning	MAT	Ortega-Bueno et al. [152]	linguistics	20,000	F-score: 0.785	Test dataset	Binary
	STD	Chen and Soo [35]	HCF, word2vec	504,118	Accuracy: 0.95	10-fold C.V	Binary
	MAT	Ortega-Bueno et al. [154]	linguistics	30,000	Accuracy: 0.82	N.A	Binary

The bold entries show the best performing results among the various state-of-the-arts using the same approach and dataset category.

7.4.1 Supervised Approaches. In supervised approaches STD, Twitter dataset, and LTD are used for humor recognition tasks.

STD: Mihalcea and Strapparava [130] considered humor recognition as a classification task to determine whether a text contains humor or not. The humorous samples include one-liners, and the non-humorous samples include Reuters titles, proverbs, and BNC sentences. They extracted humor-specific stylistic (i.e., alliteration, antonymy, and adult slang) and content-based features. They applied SVM and NB classifiers. They observed that identifying more sophisticated humor-specific features, such as semantic oppositions and ambiguity, are important for humor recognition. Mihalcea and Pulman [128] further introduced features such as human-centeredness and negative polarity. They applied SVM and NB classifiers. They observed that serious and humorous texts can be separated at the linguistic level and considered human-centeredness and negative orientation as two important characteristics.

Yang et al. [224] considered humorous language through humor recognition and humor anchor extraction. For humor generation, they considered semantic structure-based features, such

as incongruity, ambiguity, interpersonal effect, and phonetic style. They considered humorous samples from *pun of the day* and *16,000 one-liners* datasets, and non-humorous samples from news and proverbs. They applied RF classifier and proposed a simple and effective Maximal Decrement method for automatic extraction of anchors. Shahaf et al. [193] recognize humor in cartoon captions using the dataset from *The New Yorker* caption contest.[46] They considered features such as sentiment, taking expert advice, perplexity, readability, locations, and third-person and proper nouns. The expert advice is taken from the winners of the contest to capture their suggestions, such as the usage of monosyllabic, common, and simple words. Using their advice, readability is measured using *reading ease* [59] and *automated readability index* [192].

Twitter Datasets: Raz [174] classified humor in Twitter, targeting comedian accounts. They considered syntactical, pattern-based, lexical, and morphological features. They also discussed different types of humor, such as irony, wordplay, self-deprecating, fantasy, and insult. Zhang and Liu [227] recognized humor in Twitter and non-Twitter platforms. They considered features such as phonetic, morpho-syntactic, lexico-semantic, pragmatics, and affective. They applied Gradient Boosted Regression Trees (GBRT) and reported it as the first attempt at humor recognition in Twitter.

Khandelwal et al. [98] used English-Hindi mixed content from tweets for humor recognition (scraped using twitterscraper[47]). They applied features such as n-grams, BOW, and content words. They considered four classifiers: SVM, NB, RF, and extra trees. They observed that code-mixed corpus can be annotated with POS tags at word level for better results in language detection.

LTD: Morales and Zhai [138] identified humor in online reviews. They considered the Yelp challenge dataset[48] in the form of reviews. They extracted features such as content, alliteration, ambiguity, and incongruity and applied NB, perceptron, and Adaboost classifiers. Their model incorporated external text sources, such as news articles and Wikipedia pages for humor identification.

Zhang et al. [225] applied subjectivity, affective polarity, and contextual knowledge features. On the same dataset as Mihalcea and Strapparava [130], they applied an RF classifier. Liu et al. [115] exploited syntactic features for humor recognition along with features from Yang et al. [224] as baseline features. They used the Mihalcea and Strapparava [130] dataset and applied an RF classifier. They concluded that style- and content-independent syntactic structures are effective for humor recognition. Beukel and Aroyo [18] considered two new features—homophones and homographs—to recognize humor. They also considered style and ambiguity features. They considered one-liners and jokes as the humorous dataset and used news headlines, English proverbs, and Wikipedia sentences as the non-humorous dataset. They applied SVM and NB classifiers, reporting comparatively better performance on both short and long humorous texts. Liu et al. [116] considered sentiment discourse relations, such as sentiment transition and sentiment conflict, as indicators for humor recognition. They also considered features from Yang et al. [224] and applied an RF classifier. They considered humorous and non-humorous samples from Reference [130], observing that sentiment association is a better representation for humor recognition, as compared to simply detecting sentiment polarity.

7.4.2 Deep Learning–based Approaches. Recently, some deep learning–based approaches have also been applied for humor recognition over STD.

STD: Chen and Soo [35] used CNNs with extensive use of filter size and filter numbers. They introduced a highway network to implement humor recognition using one-liners [130], pun of the

[46] https://www.newyorker.com/cartoons/contest.
[47] https://github.com/taspinar/twitterscraper.
[48] https://www.yelp.com/dataset/challenge.

Table 7. A Summary of the Existing Literature on Simile Detection

Approach	Dataset	Literature	Feature	Dataset size	Eval. res.	Val. appr.	Classfication
Supervised	LTD	Niculae and Danescu-Niculescu-Mizil [148]	vehicle specificity, vehi-cle imageability	816	AUC: 0.94	5-fold C.V, Test dataset	Binary
	MAT	Qadir et al. [166]	lexical, semantic, senti-ments	2,805	F-score: 0.60	10-fold C.V	Binary
		Qadir et al. [167]	syntactic, structures	641	MRR: 0.41	N.A	Binary
Rule-Based	LTD	Hao and Veale [69]	patterns-based	35,355	F-score: 0.88	Test dataset	Binary
Deep-Learning	HAT	Manjusha and Raseek [121]	emotion, sentiment, punc-tuations	2,200	AUC: 0.94	Test dataset	Multi-class
	LTD	Liu et al. [114]	BOW, word embedding	11.3k	F-score: 0.86	5-fold C.V	Binary

day [224], and jokes dataset. They considered human centric features (HCF) from Reference [224] and also word2vec features. Sane et al. [188] considered humor recognition in Hindi-English code-mixed tweets for humor recognition. They considered two pre-trained embedding models, CNNs, and bi-LSTMs with and without attention.

7.4.3 Multilingual Studies for Humor Recognition. Ortega-Bueno et al. [152] considered Spanish social media for humor recognition. They considered both attention-based RNNs and LSTMs, and linguistics features (i.e., stylistic, structural, and affective). An LSTM is used to obtain long-term dependencies, and attention (pre- and post-level) layers are used to increase the effectiveness to classify a tweet as humorous or not. Ermilov et al. [49] recognized humor in Russian datasets containing novels, news headlines, and proverbs. They considered lexical, BOW, and structural features and used an SVM classifier. Ortega-Bueno et al. [154] applied a Bidirectional Gated Recurrent Unit (BiGRU) network followed by an attention layer and another BiGRU and considered linguistic features for humor recognition in Spanish tweets.

The authors in References [174, 193] do not use any evaluation metrics; rather, they show some analysis results. However, the discussion by Raz [174] regarding taxonomies of humor can be useful to detect varied categories of humor, and the work by Shahaf et al. [193] can be useful for multimodal platforms, such as Instagram, where captions are used in images mainly to show humorous effect.

7.5 Simile Detection Approaches

Simile and *metaphor* are comparison-based categories of figurative language. They differ from each other with respect to the connecting words, such as *"as," "like,"* and *"than."* Simile uses these connecting words to connect two different entities, whereas metaphor generally avoids the explicit usage of such connecting words. Compared to work on sarcasm and irony, simile detection works are relatively few. Table 7 presents a summary of the existing literature on simile detection.

7.5.1 Supervised Approaches. Supervised approaches are mainly used over LTD and MAT for simile detection.

LTD: Fishelov [58] states that simile provides positive and negative orientation of sentiment for an entity. Veale and Hao [212] extracted topical world knowledge from the Web regarding simile and metaphor and generated 63,935 unique adjective-noun associations through WordNet [132]. Niculae and Danescu-Niculescu-Mizil [148] collected a simile dataset from Amazon product

reviews and determined the figurative comparisons. They also emphasized that domain knowledge is essential for simile identification.

MAT: Qadir et al. [166] proposed the extraction of affective polarity (positive, negative, or neutral) from similes based on component phrases, where affective polarity is related to the state of the "topic (tenor)" in a simile. The authors used tweets to recognize similes and considered lexical (unigrams, simile components, paired components, and explicit properties associated with vehicle), semantic (hypernym class, perception verb), and sentiment (component, explicit property, simile connotation) features and applied SVM for classification. Later on, Qadir et al. [167] inferred implicit properties in open similes. They collected similes from Twitter and recognized noun and verb phrases by applying POS taggers and reported the MRR metric.

7.5.2 Rule-based Approaches. Rule-based approaches are mainly applied over LTD for simile detection based on distinct patterns.

LTD: Hao and Veale [69] proposed an algorithm to differentiate ironic similes from non-ironic similes. They identified different ironic similes from the Web and extracted different rule snippets, such as "as * as *" and "about as * as *."

7.5.3 Deep Learning–based Approaches. Recently, deep learning–based approaches have been proposed for simile detection over HAT.

HAT: Manjusha and Raseek [121] considered similes as composed of other figurative language categories, such as sarcasm, irony, and humor. They extracted sentiment, punctuation, and emotion-based features and applied CNN, SVM, DT, KNN, and Gaussian Naive Bayes (GNB) for simile detection.

7.5.4 Multilingual Studies for Simile Detection. Liu et al. [114] consider simile detection in Chinese, providing a corpus[49] of sentences. They proposed a neural learning framework over three tasks: (i) simile classification, (ii) simile component extraction to figure out "tenor" or "vehicle" in a sentence, and (iii) language modeling for predicting neighboring words.

Several works [58, 140, 212] do not provide any evaluation via metrics, but rather they show some analysis results. For example, the approach by Fishelov [58] can be used for developing sentiment analysis systems, where a classification task could be to identify positive, negative, and neutral similes. Similarly, the work in Veale and Hao [212] can be used to detect patterns from simile instances using tag pairs (e.g., adjective-noun pairs). Further, Mpouli [140] proposed a framework to annotate similes in literary texts using deep semantic and syntactic characteristics, which can be useful for simile classification tasks using machine learning techniques.

7.6 Metaphor Detection Approaches

Like simile, only relatively few works exist for the computational detection of metaphor. Moreover, application of ML techniques for metaphor detection is rare, in comparison to sarcasm and irony detection. Table 8 presents a summary of the existing literature on metaphor detection.

7.6.1 Supervised Approaches. Supervised approaches are mainly used over LTD for metaphor detection.

LTD: Shutova et al. [194] proposed an approach for automatic metaphor identification in unrestricted texts, where noun and verb clustering is applied to capture metaphorical expressions. Mohler et al. [136] detected linguistic metaphors and applied semantic similarity and compared with a set of known metaphors for a sentence. Bracewell et al. [26] considered a method in which a semantic signature is constructed for a target. Dunn [46] measured the metaphoric score as a

[49] https://github.com/cnunlp/Chinese-Simile-Recognition-Dataset.

Table 8. A Summary of the Existing Literature on Metaphor Detection

Approach	Dataset	Literature	Features	Dataset size	Eval. res.	Val. appr.	Classfication
Supervised	LTD	**Mohler et al. [137]**	**semantic pattern**	**55,895**	**F-score: 0.85**	**10-fold C.V**	**Multi-class**
		Dunn [46]	predictive features	6,893	F-score: 0.63	N.A	Binary
		Jang et al. [83]	global contextual, local contextual	2,670	F-score: 0.79	10-fold C.V	Binary
		Jang et al. [81]	topic transition	2,670	F-score: 0.81	10-fold C.V	Binary
		Mosolova et al. [139]	unigram, POS	117	F-score: 0.75	Test dataset	Binary
Semi-Supervised	LTD	Jang et al. [82]	unigram, frame	2,670	F-score: 0.82	10-fold C.V	Binary
Unsupervised	LTD	LTD Pramanick and Mitra [164]	AN pairs	1,968	Accuracy: 0.72	Test dataset	Binary
Deep-Learning	LTD	**Wu et al. [222]**	**POS, word cluster**	**28,322**	**F-score: 0.67**	**Test dataset**	**Binary**
		Pramanick et al. [163]	token, lemma	N.A	F-score: 0.67	10-fold C.V	Binary
		Swarnkar and Singh [199]	contrast	1,17,920	F-score: 0.60	Test dataset	Binary

The bold entries show the best performing results among the various state-of-the-arts using the same approach and dataset category.

scalar value between 0 and 1. The author used the Vrije Universiteit (VU) Amsterdam metaphor corpus[50] for binary classification.

Jang et al. [83] proposed a novel approach in which global contextual features, such as semantic category, topic distribution, and lexical chain, are introduced. In addition, local contextual features, semantic features, and grammatical dependencies are also considered to detect metaphors. They applied a logistic regression classifier. Jang et al. [81] proposed an approach in which sentence-level topic transitions are considered. They applied topic transition–based features, such as target sentence topic, topic difference, topic similarity, topic transition, and topic transition similarity. They applied an SVM classifier and performed 10-fold cross validation. Mosolova et al. [139] proposed metaphor detection using Conditional Random Fields (CRF).

7.6.2 Semi-Supervised Approach. **LTD**: Jang et al. [82] applied a semi-supervised bootstrapping approach to construct a metaphor frame on an unlabeled corpus.

7.6.3 Unsupervised Approach. **LTD**: Pramanick and Mitra [164] applied k-means clustering to detect metaphor. They considered features from Adjective-Noun (AN) pairs to classify instances into two disjoint classes. They used the dataset from Tsvetkov et al. [209], which provides a corpus of AN pairs.

7.6.4 Deep Learning–based Approaches. **LTD**: Pramanick et al. [163] detected metaphor at the token level on VU Amsterdam metaphor corpus. They considered features, such as applied token, lemma, and POS, and applied Condition Random Fields (CRF) and bi-directional LSTMs. Similarly, Wu et al. [222] proposed a combination of CNNs and LSTMs to detect metaphors to obtain contextual information. They also used the VU Amsterdam metaphor corpus. They represented sentences

[50] http://www.vismet.org/metcor/documentation/home.html.

at both local and long distance and considered POS and word cluster–based features. Swarnkar and Singh [199] proposed an LSTM-based contrast network approach. They used the VU Amsterdam metaphor corpus and used contrast features generated from pre-trained word embeddings.

7.6.5 Multilingual Studies for Metaphor Detection. Mohler et al. [137] applied annotations in four languages—English, Spanish, Russian, and Farsi. They proposed a four-tuple for each metaphor annotation—namely, "source," "target," "relation," and "metaphoricity." Using these tuples, a semantic patterns set is derived. Dunn et al. [47] presented a language-independent ensemble-based approach to identify metaphors in English, Spanish, Russian, and Farsi. Their system's architecture allows easy integration of new metaphor identification schemes and achieves significantly better results over multiple languages. Similarly, some other articles [11, 53, 61, 71, 95, 127, 146, 147, 212] discuss metaphor detection along with other figurative language categories, such as sarcasm, irony, and simile.

Other related work includes, for example, Reference [194], which can be extended using an unsupervised clustering approach, as they used noun and verb clustering to collect metaphoric expression. Similarly, the approach in Mohler et al. [136] can be used for detecting metaphoric patterns in a text using an NN model. Finally, the work in Bracewell et al. [26] can be used as seed metaphoric word in an expression for detecting metaphor.

7.7 Hyperbole Detection Approaches

Hyperbole plays an important role in sarcasm and irony detection. The presence of hyperbole puts an extra emphasis within the text to draw the attention of the reader. Liebrecht et al. [111] considered hyperbole as a sign of sarcastic utterances and argued that hyperbole can constitute intensifiers (adverbs, adjectives), exclamation marks, or a combination of both. Sarcasm is easier to identify as sarcastic or non-sarcastic in the presence of hyperbole. For example, the hyperbolic phrase "fantastic weather" is easier to recognize as sarcastic text instead of the non-hyperbolic phrase "the weather is good" [106]. Bharti et al. [20] found that the use of intensifiers, punctuations, interjections, and quotes in textual data are the markers of hyperbole. Bamman and Smith [8] collected a list of 50 intensifiers from Wikipedia[51] and validated their usage in hyperbole-related tweets. Lunando and Purwarianti [120] found that the presence of hyperbole in utterances makes the sarcasm detection task easier when compared to utterances without hyperbole. Tungthamthiti et al. [210] considered hyperbolic (punctuation) features to figure out the contradiction among situation and sentiment. Bharti et al. [21] considered hyperbole-based features for classification.

The works above are mainly on sarcasm and irony detection in the presence of hyperbole. Recently, Troiano et al. [207] proposed the first direct attempt towards computational detection of hyperbole. They propose a manually annotated corpus named "HYPO," which includes exaggeration (i.e., hyperbole) and non-exaggeration instances. They include qualitative and quantitative features, such as imageability, unexpectedness, polarity, subjectivity, and emotional intensity. Imageability describes the degree to which the mental image of a word is captured, and it is extracted with the help of imageability ratings available in Medical Research Council (MRC) psycholinguistic database from Tsvetkov et al. [209]. They further applied LR, KNN, DT, NB, and LDA for learning.

8 FIGURATIVE LANGUAGE: COMPUTATIONAL DIFFERENCES AND COMMONALITIES

This section presents the basic computational differences and commonalities among the various categories of figurative language.

[51] https://en.wikipedia.org/wiki/intensifier.

8.1 Computational Differences: Sarcasm versus Irony

Although sarcasm and irony are very similar in nature and both are generally used interchangeably, sarcasm is a special case of irony [85]. Irony is categorized as *verbal*, *situational*, and *dramatic*, whereas sarcasm is a *verbal* form of irony [57]. As a result, most of the existing approaches for irony detection are almost similar to the sarcasm detection approaches, including the datasets, feature-extraction process, and detection mechanism. Sarcasm is more a aggressive, offensive, and less subtle form of irony. It contains more aggregation and aims to express contempt or ridicule. However, irony is considered as sharp and non-offensive [89, 198]. Recently, some research works have considered the differences between sarcasm and irony that are discussed in the following subsection.

8.1.1 Supervised Approaches. Mostly, supervised techniques are applied over HAT and LTD datasets for sarcasm versus irony detection.

HAT: Wang [216] applied quantitative sentiment analysis and qualitative content analysis over Twitter data to study similarities and differences between sarcasm and irony. They considered *aggressiveness* as the distinguishing factor between sarcasm and irony. They mention that sarcasm contains more aggregation as compared to irony. During quantitative analysis, they also observed that sarcastic texts are more positive as compared to ironic ones. Due to the pragmatic insincerity and the divergence between what the users intend to mean and their presented expression, the aggressive intention of sarcasm is often expressed using more positive words [216]. However, in terms of qualitative content analysis, the authors noticed that sarcastic tweets usually consider a specific target. They also assume that ironic texts are written more for generic events, whereas sarcasm has specific targets.

Khokhlova et al. [100] also consider linguistic differences and analyzed tweets for sarcasm and irony. They considered eight corpora that are labeled using the #sarcasm and #irony hashtags. They observe that ironic texts use more proper names as compared to sarcastic texts. In terms of parts of speech, ironic texts contain more nouns, whereas other parts of speech are more frequently found in sarcastic texts. Like Wang [216], Khokhlova et al. [100] also found that sarcastic texts are more positive than ironic texts. Furthermore, they observe that the number of hashtags in ironic tweets are more as compared to sarcastic tweets. Ironic texts are also more structured, while sarcasm texts are more rhetorical in structure. They mention that people use different topics of their interest in irony, whereas they consider usual concepts such as "drinking" and "pastime" in sarcastic texts.

Ling and Klinger [112] considered "figurative-specific," "sentiment," and "syntactic" features and noticed that sarcastic tweets contain more positive words than ironic tweets; ironic tweets contain more tokens than sarcastic tweets; sarcastic tweets contain more positive words as compared to ironic tweets. Moreover, @usernames mentions are found more in sarcasm utterances as compared to irony tweets, considering the fact that sarcasm generally targets someone. They considered "figurative-specific," "sentiment," and "syntactic" features.

Sulis et al. [198] considered the three hashtags #sarcasm, #irony, and #not and noticed that #not is a negative indicator of sarcasm. They analyzed structural and affective features in tweets for binary classification tasks, such as #sarcasm versus #irony, #irony versus #not, and #sarcasm versus #not. Based on sentiment polarity values, they noticed that positive emotional words are found more in sarcasm and #not tweets, as compared to irony tweets. They noticed that sentiment and affective features are important for #irony versus #sarcasm task, whereas structural and sentiment analysis features are good in case of #irony versus #not. They reported that a cross-language study of sarcasm- and irony-related markers could be an interesting task for future research. They also point out that the investigation of educational and socio-demographic background of irony and sarcasm users is necessary. Finally, they proposed that investigating sarcasm versus irony tweets by excluding explicit hashtags could be interesting future work.

LTD: Joshi et al. [89] investigated sarcasm versus irony classification over LTD. They considered three binary classification tasks (i) sarcasm versus irony, (ii) sarcasm versus philosophy, and (iii) sarcasm versus philosophy using unigrams, pragmatic, and implicit and explicit sentiment features. Their dataset consists of book snippets that are annotated as sarcasm, irony, and philosophy. They revealed that sarcastic utterances include more ridicule and are more target-specific in comparison to irony.

8.2 Computational Differences: Humor versus Irony

Mostly, supervised approaches have been used for humor versus irony task over HAT.

HAT: *Irony* is considered as one of the taxonomy of *humor* [70, 174]. Reyes et al. [179] considered humor and irony detection tasks. They considered four pattern-based feature sets—ambiguity, polarity, unexpectedness, and emotional scenarios—and made an evaluation to judge representativeness and relevance. They noticed that no single feature is sufficient to discriminate between *humor* and *irony*. However, all features together provide a useful linguistic inventory to detect such figurative devices.

Barbieri and Saggion [12] considered "unexpectedness" and "incongruity" as key characteristics for both *irony* and *humor*. They consider frequency, written-spoken, intensity, structure, sentiment, synonyms, and ambiguity features. They find that the proposed features are not enough to discriminate *irony* and *humor,* because both categories have their own particular characteristics. Moreover, Gibbs et al. [63] stated that *irony* and *humor* are related to each other, and both are found in spoken as well as written language.

8.3 Computational Commonalities: Simile and Irony

Computational commonality studies between simile and irony have mainly used supervised approaches over LTD.

LTD: Hao and Veale [69] conducted a very large corpus analysis of web-harvested *similes* and identified the most interesting characteristics of ironic comparisons. They construct ironic similes using patterns such as "as * as *" and "about as * as *" to extract snippets and provide an empirical evaluation for separating ironic from non-ironic similes.

8.4 Computational Commonalities: Satire and Irony

The articles related to computational commonalities between satire and irony are based on supervised approaches.

LTD: Ravi and Ravi [171] consider one ironic and two satirical datasets and applied linguistic, semantic, psychological, and unigrams features. They identified some commonalities between *satire* and *irony* using features such as affective process (negative emotion), personal concern (leisure), biological process (body and sexual), perception (see), informal language (swear), social process (male), cognitive process (certain), and psycho-linguistic (concreteness and imageability). They considered document-term matrix, LIWC, and Tool for the Automatic Analysis of Lexical Sophistication (TAALES[52]) for extracting these features and observed that both satire and irony share common characteristics.

8.5 Computational Commonalities: Hyperbole and Sarcasm

The studies of commonalities between hyperbole and sarcasm have mainly used supervised approaches over HAT.

[52] https://www.linguisticanalysistools.org/taales.html.

HAT: Bharti et al. [20, 21] noticed that the presence of hyperbole in the form of interjections, intensifiers, quotes, and punctuations are markers of *sarcasm*, and they considered hyperbole as an important indicator for sarcasm detection. Similarly, Kunneman et al. [106] considered *hyperbole* as one of the linguistic markers to detect sarcasm. They reported that hyperbole-related exclamations and intensifiers are among the most predictive features.

9 SHARED TASKS RELATED TO FIGURATIVE LANGUAGE

In this section, we present a review of the shared tasks related to FL detection approaches. These shared tasks allow comparative evaluation of more than one approach among participating teams on a common dataset [85]. Ghosh et al. [61] described a shared task from "Sem-Eval 2015 task 11," in which 15 teams participated to perform sentiment analysis on some of the figurative language categories—sarcasm, irony, and metaphor. They considered figurative language consisting of Twitter hashtags #sarcasm, #irony, and #metaphor. The dataset consists of 8K tweets for training and 4K tweets for testing, with the aim to classify sentiment on a scale of 11 points (−5 to +5) to determine different sentiment polarity scores. The results were evaluated on the basis of the MSE value. The teams used affective resources like SenticNet for polarity scores and considered character n-grams, POS tags, and lexical features. Most of the teams, such as `elirf` and `LLT_PolyU`, performed well on sarcasm and irony tweets only. The `elirf` team considered character n-grams and a bag-of-words model and applied SVM, whereas the `LLT_PolyU` team considered a semi-supervised approach and included word-level sentiment scores and dependency labels as features. However, the `ClaC` team performed best for the metaphor category and was announced as the winning team. They used four lexica, out of which, one was automatically generated and three were manually crafted. They also considered term frequencies, POS tags, and emoticons as features.

Hee et al. [74] considered two tasks (A and B) related to irony at "SemEval-2018 task-3." Task-A was to determine whether a tweet is ironic or not, and task-B was to determine either the type of irony (i.e., verbal/situational) or the tweet being non-ironic. As a shared task, 43 teams participated for task-A and 31 teams for task-B, including Farías et al. [52], Hernández-Farías et al. [76], Pamungkas and Patti [155], Peng et al. [159], Vu et al. [213]. Their approaches considered features ranging from hand-crafted (e.g., sentiment, syntactic, and semantic) to character and word embeddings. For both tasks, the training dataset consists of 3,834 tweets, and the test dataset consists of 784 tweets. ML classifiers, such as SVM, RF, ME, NB, and deep learning–based techniques, such as CNNs, RNNs, and bi-LSTMs, were used. The `THU_NGN` team reported an F-score of 0.71 on task-A. They used densely connected LSTMs, based on a pre-trained word-embedding architecture, and considered sentiment- and syntactic-based features. However, the `UCDCC` team reported an F-score of 0.51 on task-B and employed a *siamese* architecture [27].

Potash et al. [162] presented a shared task from Sem-Eval 2017 task 6 on humor recognition. A total of eight teams and 19 systems participated for two sub-tasks—A and B, with a total of 12,734 tweets spanning 112 hashtags. The participating teams used incongruity, ambiguity, and stylistic features. Sub-task A was to determine which tweet within a pair was funnier. For sub-task B, teams were asked to determine the labels directly by providing prediction files in which tweets are ranked on the basis of funny content. For sub-task A, participating teams used either feature-based systems or neural network–based systems. Top teams for sub-task A preferred neural network–based systems. The `HumorHawk` team reported the highest accuracy of 0.675. They used an ensemble system that utilized predictions from both feature-based and neural network–based systems. For sub-task B, `Duluth` reported the best result in terms of edit distance metric, which measures the distance between actual and predicted labels. They applied the same approach as `SVNIT` and `QUB` teams but used the output of a language model to rank the tweets, as opposed to labels provided by a classifier.

Castro et al. [32] proposed the Humor Analysis based on Human Annotation (HAHA) shared task from IberEval 2018 workshop. The HAHA task consists of two sub-tasks related to automatic humor recognition in Spanish. The first sub-task was used for humor recognition, whereas the second sub-task aims at funniness score prediction. The proposed systems were mainly based on neural network and machine learning techniques. The INGEOTEC team performed best for both sub-tasks. Interestingly, they proposed an evolutionary algorithm-based EvoDAG system that produced the best result and obtained an F-score of 0.79 for the first sub-task. However, a regression model based on kernel ridge regression by the same team performed best for the second sub-task.

10 DISCUSSION AND OPEN CHALLENGES

Based on the in-depth review of existing literature on different figurative language detection approaches, we observe that most of the researchers have considered supervised explicit markers like hashtags to generate datasets for different figurative language detection tasks. Similarly, most of the approaches have considered only textual data appearing in tweets. However, the presence of other forms of data (e.g., images, videos) could be used to express figurative language categories. The existence of multi-lingual texts (mainly using regional languages) is another issue for developing figurative language detection systems. Although English texts are the most popular choice for researchers, changes in geographical location generally lead to use of a new language, or at least the usage of regional slangs. Therefore, development of language-independent figurative language detection systems is a challenging and promising research task.

Below, we summarize some of the open challenges related to figurative language detection.

- *Development of multimodal systems*: Social media data are a mixture of texts, images, audio, and video, and there is a possibility of sarcastic and ironic expressions involving each of them. However, most of the authors have considered only textual data for figurative language detection. As a result, multimodal systems for figurative language detection are very rare. Recently, Schifanella et al. [191] proposed the first multimodal system for sarcasm detection on Twitter, Instagram, and Tumblr, and Das and Clark [40] on Flickr images. However, this is clearly an important future direction.
- *Cross-domain detection systems*: Since the availability of labeled data for training supervised machine learning models is one of the basic requirements, and there is a scarcity of labeled data in the field of figurative language detection, the exploration of cross-domain machine learning techniques, such as *transfer learning*, is a promising direction of research. In transfer learning, a classification model is trained over the dataset of one domain to solve the classification problem of another related domain.
- *Stream-data analysis*: Most of the existing approaches for figurative language detection analyze data offline. Therefore, devising computationally efficient approaches with rigorous pre-processing and filtering techniques is a must for efficient processing of streaming data and constitutes a promising future direction of research. Bharti et al. [21] recently attempted sarcastic sentiment detection in Twitter stream data, but it is still a challenging task and needs further work.
- *Deep learning and ensemble approaches*: Deep learning has been successful in dealing with textual data, and therefore exploring the full potential of deep learning techniques for computational detection of figurative language in texts, images, and other forms of data is a promising direction of research. An accurate semantic representation of an instance and extraction of definitive information are important steps to retrieve the exact meaning from an instance, especially in case of figurative language. For example, in References [5, 35, 45, 62, 73, 161, 189, 191, 199, 226], the authors used deep learning–based semantic

modeling to address varied categories of figurative language. Application of classifier ensemble techniques is another research direction, which has been applied infrequently for figurative language detection [47, 56, 117, 206].

- *Authentic benchmarks generation*: As discussed earlier, hashtag-based tweet annotation is one of the methods to generate annotated Twitter datasets. In this approach, tweets are labeled based on the presence of certain hashtags such as #sarcasm, #not, #irony, and #humor that are mentioned by the authors in their tweets. However, it is quite possible that normal tweets are wrongly tagged as figurative language tweets using these tags, either intentionally or by mistake. However, figurative language tweets may not have been tagged using any of these hashtags. Therefore, manual verification of labels by domain experts to create authentic benchmarks is an important, though time-consuming, task.
- *Macaronic language detection*: Figurative language detection in English is quite popular among researchers, though some of the researchers have also focused on other languages. Macaronic language is a mixing of languages in one text, often for humor,[53] which is generally used by non-English speakers, and it is very common in online social media. For example, Hinglish[54] is a macaronic language that uses both Hindi and English words, e.g., "*iskool*" for school. Therefore, the presence of macaronic language in a dataset adds another dimension of challenge, and figurative language detection in such datasets is a challenging task.
- *Comparison and commonalities between figurative language categories*: Different categories of figurative language have many commonalities, such as *sarcasm* and *irony*, *humor* and *irony*, and *sarcasm* and *humor*. For example, *sarcasm* and *irony* are interchangeably used by some of the researchers. However, there are some works that discriminate sarcasm and irony [85, 100, 112, 198, 216] on the basis of *aggressiveness*, *ridicule*, *target*, and *presence of positive words* factors. Similarly, *unexpectedness* and *incongruity* are key factors for irony and humor [12]. Further, *hyperbole* in the form of *exaggeration*, such as intensifier and interjection, are found in sarcasm utterances [20, 21, 106]. Due to many such commonalities between the figurative language categories, their computational detection needs more fine-grained approaches and is a promising area of research.

11 CONCLUSION

We have presented an in-depth survey of computational detection methods for figurative language in various online data sources, such as tweets, reviews, blogs, and e-news. Since the presence of figurative language greatly impacts the actual polarity and interpretation of the sentiment bearing words, their computational detection is vital for the development of better systems for sentiment analysis, user profiling, recommendations, brand endorsement, and product campaigning [50, 93, 94, 125]. In this review, we have considered seven major categories of figurative language—*sarcasm*, *irony*, *satire*, *humor*, *simile*, *metaphor*, and *hyperbole*. Starting from their basic definitions and historical evolution, we presented details about their characteristic features, datasets used, and various state-of-the-art computational detection techniques. Our discussion includes challenges and future directions of research for each figurative language category. We hope that this survey is a useful resource for researchers, especially for new researchers who plan to start their research career in the field of text data analytics, computational linguistic, natural language processing, social computing, or figurative language detection using computational techniques.

[53] https://en.wikipedia.org/wiki/Category:Macaronic_language.
[54] https://en.wikipedia.org/wiki/Hinglish.

REFERENCES

[1] Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics–Student Research Workshop (ACL-SRW'16)*. ACL, 107–113.

[2] Muhammad Abulaish and Ashraf Kamal. 2018. Self-deprecating sarcasm detection: An amalgamation of rule-based and machine learning approach. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'18)*. IEEE, 574–579.

[3] Ameeta Agrawal and Aijun An. 2018. Affective representations for sarcasm detection. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. ACM, 1029–1032.

[4] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Hidden Markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*. AAAI, 3–10.

[5] Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of the 20th Special Interest Group on Natural Language Learning Conference on Computational Natural Language Learning (SIGNLL-CoNLL'16)*. ACL, 167–177.

[6] Salvatore Attardo. 2010. *Linguistic Theories of Humor*. Vol. 1. Walter de Gruyter.

[7] Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor-Int. J. Humor Res.* 4, 3–4 (1991), 293–348.

[8] David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on Twitter. In *Proceedings of the 9th International Association for the Advancement of Artificial Intelligence Conference on Web and Social Media (ICWSM'15)*. Citeseer, 574–577.

[9] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. Do we criticise (and laugh) in the same way? Automatic detection of multi-lingual satirical news in Twitter. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*. 1215–1221.

[10] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. Is this tweet satirical? A computational approach for satire detection in Spanish. *Proc. Lenguaje Nat.* 55 (2015), 135–142.

[11] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. UPF-taln: Semeval 2015 tasks 10 and 11 sentiment analysis of literal and figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 704–708.

[12] Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in Twitter. In *Proceedings of the 5th International Conference on Computational Creativity (ICCC'14)*.

[13] Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics (SRW-EACL'14)*. ACL, 56–64.

[14] Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in Twitter: Feature analysis and evaluation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 4258–4264.

[15] Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in Twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-ACL'14)*. ACL, 50–58.

[16] Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 sentiment polarity classification task. In *Proceedings of the 4th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*. 50–57.

[17] Monroe C. Beardsley. 1981. *Aesthetics, Problems in the Philosophy of Criticism* (2nd ed.). Hackett Publishing Company, Inc., Indianapolis, IN.

[18] Sven V. D. Beukel and Lora Aroyo. 2018. Homonym detection for humor recognition in short text. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-ACL'18)*. ACL, 286–291.

[19] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2017. Harnessing online news for sarcasm detection in Hindi tweets. In *Proceedings of the 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI'17)*, B. Uma Shankar, Kuntal Ghosh, Deba Prasad Mandal, Shubhra Sankar Ray, David Zhang, and Sankar K. Pal (Eds.). Springer, 679–686.

[20] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena Jena. 2015. Parsing-based sarcasm sentiment recognition in Twitter data. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15)*. IEEE, 1373–1380.

[21] Santosh Kumar Bharti, Bakhtyar Vachha, Ramkrushna Pradhan, Korra Sathya Babu, and Sanjay Kumar Jena. 2016. Sarcastic sentiment detection in tweets streamed in real time: A big data approach. *Dig. Commun. Netw.* 2, 3 (July 2016), 108–121.

[22] Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'13)*. 4158–4162.
[23] Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Opinion mining in Twitter how to make use of sarcasm to enhance sentiment analysis. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15)*. IEEE, 1594–1597.
[24] Mondher Bouazizi and Tomoaki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on Twitter. *IEEE Access* 4 (Sept. 2016), 5477–5488.
[25] Andrea Bowes and Albert Katz. 2011. When sarcasm stings. *Disc. Proc.* 48, 4 (2011), 215–236.
[26] David B. Bracewell, Marc T. Tomlinson, and Michael Mohler. 2013. Determining the conceptual space of metaphoric expressions. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13)*. Springer, 487–500.
[27] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a "siamese" time delay neural network. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'94)*. 737–744.
[28] Marc Brysbaert, Amy B. Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Meth.* 46, 3 (2014), 904–911.
[29] Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP'09)*. ACL and AFNLP, 161–164.
[30] Arnie Cann, Lawrence G. Calhoun, and Janet S. Banks. 1997. On the role of humor appreciation in interpersonal attraction: It's no joking matter. *Humor-Int. J. Humor Res.* 10, 1 (1997), 77–90.
[31] Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! It's so easy;-). In *Proceedings of the 1st International Workshop on Topic-sentiment Analysis for Mass Opinion (TSA'09)*. ACM, 53–56.
[32] Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018. Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018. In *Proceedings of the IberEval Workshop*. 187–194.
[33] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3 (2011), 27.
[34] Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets. *Eng. Appl. Artif. Intell.* 51 (Mar. 2016), 50–57.
[35] Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 16th Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), (NAACL-HLT'18)*. ACL, 113–117.
[36] Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *J. Exper. Psychol.: Gen.* 113, 1 (1984), 121–126.
[37] Conal Condren. 2014. Satire. In *Encyclopedia of Humor Studies*, Salvatore Attardo (Ed.). SAGE Publications.
[38] Marlena A. Creusere. 1999. Theories of adults' understanding and use of irony and sarcasm: Applications to and evidence from research with children. *Dev. Rev.* 19, 2 (1999), 213–262.
[39] Dipto Das and Anthony J. Clark. 2018. Sarcasm detection on Facebook: A supervised learning approach. In *Proceedings of the International Conference on Multimodal Interaction: Adjunct (ICMI'18)*. ACM, 3.
[40] Dipto Das and Anthony J. Clark. 2018. Sarcasm detection on Flickr using a CNN. In *Proceedings of the International Conference on Computing and Big Data (ICCBD'18)*. ACM, 56–61.
[41] Hal Daumé, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Mach. Learn.* 75, 3 (2009), 297–325.
[42] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CONLL'10)*. ACL, 107–116.
[43] Larissa A. de Freitas, Aline A. Vanin, Denise N. Hogetop, Marco N. Bochernitsan, and Renata Vieira. 2014. Pathways for irony detection in tweets. In *Proceedings of the 29th Association for Computing Machinery Symposium on Applied Computing (SAC'14)*. ACM, 628–633.
[44] Henk G. G. M. Driessen and James D. Wright. 2015. In *Anthropology of Humor*, James D. Wright (Ed.). Elsevier, 416–419.
[45] Abhijeet Dubey, Aditya Joshi, and Pushpak Bhattacharyya. 2019. Deep models for converting sarcastic utterances into their non sarcastic interpretation. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD'19)*. ACM, 289–292.

[46] Jonathan Dunn. 2014. Measuring metaphoricity. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics (ACL'14)*. ACL, 745–751.
[47] Jonathan Dunn, Jon Beltran de Heredia, Maura Burke, Lisa Gandy, Sergey Kanareykin, Oren Kapah, Matthew Taylor, Dell Hines, Ophir Frieder, David Grossman, Newton Howard, Moshe Koppel, Scott Morris, Andrew Ortony, and Shlomo Argamon. 2014. Language-independent ensemble approaches to metaphor identification. In *Proceedings of the 28th Association for the Advancement of Artificial Intelligence Workshop on Cognitive Computing for Augmented Human Intelligence (AAAI'14)*. AAAI.
[48] Sayandip Dutta and Anit Chakraborty. 2019. A deep learning–inspired method for social media satire detection. In *Soft Computing and Signal Processing*, Jiacun Wang, G. Ram Mohana Reddy, V. Kamakshi Prasad, and V. Sivakumar Reddy (Eds.). Springer, 243–251.
[49] Anton Ermilov, Natasha Murashkina, Valeria Goryacheva, and Pavel Braslavski. 2018. Stierlitz meets SVM: Humor detection in Russian. In *Proceedings of the 7th Conference on Artificial Intelligence and Natural Language (AINL'18)*. Springer, 178–184.
[50] D. I. Hernández Farias and Paolo Rosso. 2017. Irony, sarcasm, and sentiment analysis. In *Sentiment Analysis in Social Networks*, Federico A. Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu (Eds.). Elsevier, 113–128.
[51] Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in Twitter: The role of affective content. *ACM Trans. Internet Technol.* 16, 3 (June 2016), 19.
[52] Delia Irazú Hernández Farías, Fernando Sánchez-Vega, Manuel Montes y Gómez, and Paolo Rosso. 2018. INAOE-UPV at SemEval-2018 task 3: An ensemble approach for irony detection in Twitter. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval'18)*. ACL, 594–599.
[53] Delia Irazú Hernández Farías, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2015. ValenTo: Sentiment analysis of figurative language tweets with irony and sarcasm. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 694–698.
[54] C. Fellbaum. 1998. WordNet. *The Encyclopedia of Applied Linguistics*. Wiley Online Library.
[55] Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Meeting of the Association for Computational Linguistics (ACL'13) (Volume 1: Long Papers)*. ACL, 1774–1784.
[56] Elisabetta Fersini, Federico A. Pozzi, and Enza Messina. 2015. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA'15)*. IEEE, 1–8.
[57] Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), 392–398.
[58] David Fishelov. 2007. Shall I compare thee? Simile understanding and semantic categories. *J. Lit. Sem.* 36, 1 (Apr. 2007), 71–87.
[59] Rudolph Flesch. 1948. A new readability yardstick. *J. Applied Psych.* 32, 3 (1948), 221.
[60] Northrop Frye. 1944. The nature of satire. *Univ. Toronto Quart.erly* 14, 1 (1944), 75–89.
[61] Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 470–478.
[62] Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 15th North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. ACL, 161–169.
[63] Raymond W. Gibbs, Gregory A. Bryant, and Herbert L. Colston. 2014. Where is the humor in verbal irony? *Humor* 27, 4 (2014), 575–595.
[64] Rachel Giora. 1995. On irony and negation. *Disc. Proc.* 19, 2 (1995), 239–264.
[65] Paul H. Grice. 1975. *Logic and Conversation*. Harward University Press. 41–58.
[66] Charles R. Gruner. 1997. *The Game of Humor: A Comprehensive Theory of Why We Laugh*. Transaction Publishers.
[67] Raj Kumar Gupta and Yinping Yang. 2017. CrystalNest at SemEval-2017 task 4: Using sarcasm detection for enhancing sentiment classification and quantification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. ACL, 626–633.
[68] Patrick Hanks. 2005. Similes and sets: The English preposition like. *Languages and Linguistics: Festschrift for Fr. Cermak. Charles University, Prague*.
[69] Yanfen Hao and Tony Veale. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Mach.* 20, 4 (2010), 635–650.
[70] Jennifer Hay. 1995. *Gender and Humour: Beyond a Joke*. Master's thesis. Victoria University of Wellington, Wellington, New Zealand.

[71] Cynthia V. Hee, Els Lefever, and Véronique Hoste. 2015. LT3: Sentiment analysis of figurative tweets: Piece of cake# notreally. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 684–688.
[72] Cynthia V. Hee, Els Lefever, and Véronique Hoste. 2016. Exploring the realization of irony in Twitter data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 1794–1799.
[73] Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Exploring the fine-grained analysis and automatic detection of irony on Twitter. *Lang. Res. Eval.* 52, 3 (2018), 707–731.
[74] Cynthia V. Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval'18)*. ACL, 39–50.
[75] Andrew D. Hepburn. 1875. *Manual of English Rhetoric*. Wilson, Hinkle & Company.
[76] Delia I. Hernández-Farías, Viviana Patti, and Paolo Rosso. 2018. ValenTO at SemEval-2018 task 3: Exploring the role of affective content for detecting irony in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval'18)*. ACL, 643–648.
[77] Joyce Oramel Hertzler. 1970. *Laughter: A Socio-scientific Analysis*. Exposition Press.
[78] Gilbert Highet. 1972. *The Anatomy of Satire*. Princeton University Press.
[79] Yu-Hsiang Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. Irony detection with attentive recurrent neural networks. In *Advances in Information Retrieval*, Joemon M. Jose, Claudia Hauff, Ismail Sengor Altingovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Eds.). Springer, 534–540.
[80] Michael Israel, Jennifer Riddle Harding, and Vera Tobin. 2004. On simile. *Lang. Cult. Mind* 100 (2004).
[81] Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rose. 2016. Metaphor detection with topic transition, emotion, and cognition in context. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (ACL'16) (Volume 1: Long Papers)*. ACL, 216–225.
[82] Hyeju Jang, Keith Maki, Eduard Hovy, and Carolyn Penstein Rose. 2017. Finding structure in figurative language: Metaphor detection with topic-based frames. In *Proceedings of the 18th Conference on Special Interest Group on Discourse and Dialogue (SIGDIAL'17)*. ACL, 320–330.
[83] Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rose. 2015. Metaphor detection in discourse. In *Proceedings of the 16th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'15)*. ACL, 384–392.
[84] Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. ACM, 815–824.
[85] Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.* 50, 5 (2017), 73.
[86] Aditya Joshi, Diptesh Kanojia, Pushpak Bhattacharyya, and Mark James Carman. 2017. Sarcasm suite: A browser-based engine for sarcasm detection and generation. In *Proceedings of the 31st Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI'17)*. AAAI, 5095–5096.
[87] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*. ACL, 757–762.
[88] Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from tv series "Friends." In *Proceedings of the 20th Special Interest Group on Natural Language Learning (SIGNLL'16) Conference on Computational Natural Language Learning (CoNLL'16)*. ACL, 146–155.
[89] Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark James Carman, Meghna Singh, Jaya Saraswati, and Rajita Shukla. 2016. How challenging is sarcasm versus irony classification?: An analysis from human and computational perspectives. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA'16)*. Australasian Language Technology Association (ALTA), 123–127.
[90] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*. ACL, 1006–1011.
[91] Frank Stringfellow Jr.1994. *The Meaning of Irony*. State University of New York.
[92] Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowl.-based Syst.* 69 (June 2014), 124–133.
[93] Ashraf Kamal and Muhammad Abulaish. 2019. An LSTM-based deep learning approach for detecting self-deprecating sarcasm in textual data. In *Proceedings of the 16th International Conference on Natural Language Processing (ICON'19)*. ACL, 1–10.
[94] Ashraf Kamal and Muhammad Abulaish. 2019. Self-deprecating humor detection: A machine learning approach. In *Proceedings of the 16th International Conference of the Pacific Association for Computational Linguistics (PACLING'19)*. Springer, 1–13.

[95] Maria Karanasou, Christos Doulkeridis, and Maria Halkidi. 2015. DsUniPi: An SVM-based approach for sentiment analysis of figurative language on Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 709–713.

[96] Jihen Karoui, Benamara Farah, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, Vol. 1. ACL, 262–272.

[97] Jihen Karoui, Farah B. Zitoune, Véronique Moriceaue, Nathalie Aussenac-Gilles, and Lamia H. Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*. ACL, 644–650.

[98] Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor Detection in English-Hindi Code-mixed Social Media Content: Corpus and Baseline System. arXiv preprint arXiv:1806.05513.

[99] Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2015. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-ACL'15)*. ACL, 25–30.

[100] Maria Khokhlova, Viviana Patti, and Paolo Rosso. 2016. Distinguishing between irony and sarcasm in social media texts: Linguistic observations. In *Proceedings of the International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT'16)*. IEEE, 1–6.

[101] Charles A. Knight. 2004. *The Literature of Satire*. Cambridge University Press.

[102] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. 2013. Ontology-based sentiment analysis of Twitter posts. *Expert Syst. Appl.* 40, 10 (2013), 4065–4074.

[103] Roger J. Kreuz and Gina M. Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language (ACL'07)*. ACL, 1–4.

[104] Roger J. Kreuz and Richard M. Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaph. Symb.* 10, 1 (1995), 21–31.

[105] Sachi Kumon-Nakamura, Sam Glucksberg, and Mary Brown. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. *J. Exper. Psych.: Gen.* 124, 1 (Mar. 1995), 3.

[106] Florian Kunneman, Christine Liebrecht, Margot V. Mulken, and Antal V. D. Bosch. 2015. Signaling sarcasm: From hyperbole to hashtag. *Inform. Proc. Manag.* 51, 4 (Aug. 2015), 500–509.

[107] Heather L. LaMarre, Kristen D. Landreville, and Michael A. Beam. 2009. The irony of satire: Political ideology and the motivation to see what you want to see in the *Colbert Report*. *Int. J. Press/Polit.* 14, 2 (2009), 212–231.

[108] G. Harry McLaughlin. 1969. SMOG grading—A new readability formula. *J. Read.* 12, 8 (1969), 639–646.

[109] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[110] John S. Leggitt and Raymond W. Gibbs. 2000. Emotional reactions to verbal irony. *Disc. Proc.* 29, 1 (2000), 1–24.

[111] Christine Liebrecht, Florian Kunneman, and Antal V. D. Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-ACL'13)*. ACL, 29–37.

[112] Jennifer Ling and Roman Klinger. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *Proceedings of the International Semantic Web Conference (ESWC'16)*, Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer, and Christoph Lange (Eds.). Springer, 203–216.

[113] Nick Littlestone. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.* 2, 4 (1988), 285–318.

[114] Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. ACL, 1543–1553.

[115] Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Exploiting syntactic structures for humor recognition. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*. 1875–1883.

[116] Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (ACL'18) (Short Papers)*. ACL, 586–591.

[117] Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. 2014. Sarcasm detection in social media based on imbalanced classification. In *Web-Age Information Management*, Feifei Li, Guoliang Li, Seung won Hwang, Bin Yao, and Zhenjie Zhang (Eds.). Springer International Publishing, 459–471.

[118] Joan Lucariello. 2007. Situational irony: A concept of events gone away. *Irony in Language and Thought: A Cognitive Science Reader*. Psychology Press, 467–498.

[119] Stephanie Lukin and Marilyn Walker. 2013. Really? Well. Apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language in Social Media (ACL'13)*. ACL, 30–40.
[120] Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS'13)*. IEEE, 195–198.
[121] P. D. Manjusha and C. Raseek. 2018. Convolutional neural network based simile classification system. In *Proceedings of the IEEE International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR'18)*. IEEE, 1–5.
[122] Jill Mann. 1973. *Chaucer and Medieval Estates Satire: The Literature of Social Classes and the General Prologue to the Canterbury Tales*. The University of Chicago Press.
[123] James H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc.
[124] Zachary J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Comput. Ling.* 30, 1 (2004), 23–44.
[125] Diana Maynard and Mark A. Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 4238–4243.
[126] Skye McDonald. 2007. Neuropsychological studies of sarcasm. In *Irony in Language and Thought*, Raymond W. Gibbs Jr. and Herbert L. Colston (Eds.). Psychology Press, 217–230.
[127] Sarah McGillion, Héctor M. Alonso, and Barbara Plank. 2015. CPH: Sentiment analysis of figurative language on Twitter #easypeasy #not. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 699–703.
[128] Rada Mihalcea and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the 16th Conference on the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'07)*. Springer, 337–347.
[129] Rada Mihalcea and Carlo Strapparava. 2005. Computational laughing: Automatic recognition of humorous one-liners. In *Proceedings of the Cognitive Science Conference*. Citeseer, 1513–1518.
[130] Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*. ACL, 531–538.
[131] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'13)*. 3111–3119.
[132] George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[133] Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the 13th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI'16)*. AAAI.
[134] Abhijit Mishra, Diptesh Kanojia, Abhijit Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (ACL'16)*. ACL, 1095–1104.
[135] Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*. ACL, 599–608.
[136] Michael Mohler, David Bracewell, David Hinote, and Marc Tomlinson. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the 1st Workshop on Metaphor in Natural Language Processing (Meta4NLP'13)*. ACL, 27–35.
[137] Michael Mohler, Marc Tomlinson, and Bryan Rink. 2015. Cross-lingual semantic generalization for the detection of metaphor. *Int. J. Comput. Ling. Appl.* 6, 2 (Feb. 2015), 117–140.
[138] Alex Morales and ChengXiang Zhai. 2017. Identifying humor in reviews using background text sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*. ACL, 492–501.
[139] Anna Mosolova, Ivan Bondarenko, and Vadim Fomin. 2018. Conditional random fields for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*. ACL, 121–123.
[140] Suzanne Mpouli. 2017. Annotating similes in literary texts. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA'17)*. 26–36.
[141] Douglas C. Muecke. 1982. *Irony and the Ironic*. Methuen.
[142] Shubhadeep Mukherjee and Pradip Kumar Bala. 2017. Sarcasm detection in microblogs using naïve Bayes and fuzzy clustering. *Tech. Soc.* 48 (2017), 19–27.

[143] Smaranda Muresan, Roberto Gonzalez-Ibanez, Debanjan Ghosh, and Nina Wacholder. 2016. Identification of nonliteral language in social media: A case study on sarcasm. *J. Assoc. Inform. Sci. Technol.* 67, 11 (Nov. 2016), 2725–2737.
[144] Constantine Nakassis and Jesse Snedeker. 2001. Beyond sarcasm: Intonation and context as relational cues in children's recognition of irony. In *Proceedings of the 26th Boston University Conference on Language Development (BUCLD'01)*. Cascadilla Press, Somerville, MA, 429–440.
[145] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Meeting of the Association for Computational Linguistics (ACL'11)*. ACL, 581–586.
[146] Hoang L. Nguyen and Jai E. Jung. 2016. Statistical approach for figurative sentiment analysis on social networking services: A case study on Twitter. *Multim. Tools Appl.* (Apr. 2016), 1–14.
[147] Hoang L. Nguyen, Trung D. Nguyen, and Dosam Hwang. 2015. Kelabteam: A statistical approach on figurative language sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 679–683.
[148] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. ACL, 2008–2018.
[149] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the Workshop on Making Sense of Microposts: Big Things Come in Small Packages (ESWC'11)*. 93–98.
[150] Luke De Oliveira and Alfredo L. Rodrigo. 2015. Humor detection in Yelp reviews. Retrieved on December 15, 2019 from https://cs224d.stanford.edu/reports/OliveiraLuke.pdf.
[151] Silviu Oprea and Walid Magdy. 2019. iSarcasm: A Dataset of Intended Sarcasm. arXiv preprint arXiv:1911.03123.
[152] Reynier Ortega-Bueno, Carlos E. Muniz-Cuza, José E. Medina Pagola, and Paolo Rosso. 2018. UO UPV: Deep linguistic humor detection in Spanish social media. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval'18) co-located with the 34th Conference of the Spanish Society for Natural Language Processing (SEPLN'18)*. 204–213.
[153] Reynier Ortega-Bueno, Francisco Rangel, Delia I. H. Farıas, Paolo Rosso, Manuel Montes y Gómez, and José E. Medina-Pagola. 2019. Overview of the task on irony detection in Spanish variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF'19), co-located with the 35th Conference of the Spanish Society for Natural Language Processing (SEPLN'19)*. CEUR-WS.org, 229–256.
[154] Reynier Ortega-Bueno, Paolo Rosso, and José E. Medina Pagola. 2019. UO UPV2 at HAHA 2019: BiGRU neural network informed with linguistic features for humor recognition. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF'19), co-located with the 35th Conference of the Spanish Society for Natural Language Processing (SEPLN'19)*. CEUR-WS.org, 212–221.
[155] Endang W. Pamungkas and Viviana Patti. 2018. #NonDicevoSulSerio at SemEval-2018 task 3: Exploiting emojis and affective content for irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval'18)*. ACL, 649–654.
[156] Natalie Parde and Rodney Nielsen. 2018. Detecting sarcasm is extremely easy;-. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles (SemBEaR'18)*. ACL, 21–26.
[157] Krishna Parmar, Nivid Limbasiya, and Maulik Dhamecha. 2018. Feature based composite approach for sarcasm detection using MapReduce. In *Proceedings of the 2nd International Conference on Computing Methodologies and Communication (ICCMC'18)*. IEEE, 587–591.
[158] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet:: Similarity: Measuring the relatedness of concepts. In *Demonstration Papers at Human Language Technologies: North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*. ACL, 38–41.
[159] Bo Peng, Jin Wang, and Xuejie Zhang. 2018. YNU-HPCC at SemEval-2018 task 3: Ensemble neural network models for irony detection on Twitter. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval'18)*. ACL, 622–627.
[160] John Peter. 1956. *Complaint and Satire in Early English Literature*. Clarendon.
[161] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'16)*. 1601–1612.
[162] Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6:# Hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. ACL, 49–57.
[163] Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*. ACL, 67–75.
[164] Malay Pramanick and Pabitra Mitra. 2018. Unsupervised detection of metaphorical adjective-noun pairs. In *Proceedings of the Workshop on Figurative Language Processing*. ACL, 76–80.
[165] Tomás Ptácek, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14)*. 213–223.

[166] Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2015. Learning to recognize affective polarity in similes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. ACL, 190–200.
[167] Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2016. Automatically inferring implicit properties in similes. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. ACL, 1223–1232.
[168] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on Twitter: A behavioral modeling approach. In *Proceedings of the 8th Association for Computing Machinery International Conference on Web Search and Data Mining (WSDM'15)*. ACM, 97–106.
[169] Victor Raskin. 1979. Semantic mechanisms of humor. In *Proceedings of the 5th Meeting of the Berkeley Linguistics Society*. 325–335.
[170] Victor Raskin, Christian F. Hempelmann, and Julia M. Taylor. 2009. How to understand and assess a theory: The evolution of the SSTH into the GTVH and now into the OSTH. *J. Lit. Theor.* 3, 2 (2009), 285–311.
[171] Kumar Ravi and Vadlamani Ravi. 2017. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowl.-based Syst.* 120 (2017), 15–33.
[172] Kumar Ravi and Vadlamani Ravi. 2018. Irony detection using neural network language model, psycholinguistic features, and text mining. In *Proceedings of the 12th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC'18)*. IEEE, 254–260.
[173] Raymond W. Gibbs Jr. and Herbert L. Colston. 2007. *Irony in Language and Thought: A Cognitive Science Reader*. Psychology Press.
[174] Yishay Raz. 2012. Automatic humor classification on Twitter. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Student Research Workshop (NAACL HLT'12)*. ACL, 66–70.
[175] Aishwarya N. Reganti, Tushar Maheshwari, Upendra Kumar, Amitava Das, and Rajiv Bajpai. 2016. Modeling satire in English text for automatic detection. In *Proceedings of the Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE'16)*. IEEE.
[176] Antonio Reyes and Paolo Rosso. 2011. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT'11)*. ACL, 118–124.
[177] Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Dec. Supp. Syst.* 53, 4 (May 2012), 754–760.
[178] Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: Beyond a simple case of negation. *Knowl. Inform. Syst.* 40, 3 (2014), 595–614.
[179] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng.* 74 (June 2012), 1–12.
[180] Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Lang. Resour. Eval.* 47, 1 (2013), 239–268.
[181] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra D. Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*. ACL, 704–714.
[182] Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: A French lexicon of discourse connectives. *Disc., Multidisc. Persp. Sig. Text Org.* 10 (2012).
[183] Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. ACL, 7–17.
[184] Willibald Ruch. 2001. The perception of humor. In *Emotions, Qualia, and Consciousness*, Alfred Kaszniak (Ed.). World Scientific, 410–425.
[185] Jason Rutter. 1997. *Stand-Up as Interaction: Performance and Audience in Comedy Venues*. Ph.D. Thesis. *University of Salford, UK*.
[186] María Del Pilar Salas-Zárate, Mario A. Paredes-Valverde, Miguel A. Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowl.-based Syst.* 128 (2017), 20–33.
[187] Mary Jane C. Samonte, Carl Justine T. Dollete, Paolo Mikkael M. Capanas, Maristela Louise C. Flores, and Caroline B. Soriano. 2018. Sentence-level sarcasm detection in English and Filipino tweets. In *Proceedings of the 4th International Conference on Industrial and Business Engineering (ICIBE'18)*. ACM, 181–186.
[188] Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019. Deep learning techniques for humor detection in Hindi-English code-mixed tweets. In *Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. ACL, 57–61.

[189] Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*. 3371–3380.

[190] Leon Satterfield. 1982. The ironic sign. In *Semiotics 1980*, Michael Herzfeld and Margot D. Lenbart (Eds.). Springer Science & Business Media, 467–474.

[191] Rossano Schifanella, Paloma D. Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the ACM on Multimedia Conference (ACM MM'16)*. ACM, 1136–1145.

[192] R. J. Senter and Edgar A. Smith. 1967. *Automated Readability Index*. Technical Report. University of Cincinnati, Ohio, USA.

[193] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. ACM, 1065–1074.

[194] Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. 1002–1010.

[195] Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. In *Radical Pragmatics*, Peter Cole (Ed.). Elsevier, 295–318.

[196] Andreas Stöckl. 2018. Detecting Satire in the News with Machine Learning. arXiv preprint arXiv:1810.00593.

[197] Marco Stranisci, Cristina Bosco, Delia I. H. Farìas, and Viviana Patti. 2016. Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 2892–2899.

[198] Emilio Sulis, Delia I. H. Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm, and #not. *Knowl.-based Syst.* 108 (May 2016), 132–143.

[199] Krishnkant Swarnkar and Anil Kumar Singh. 2018. Di-LSTM contrast: A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*. ACL, 115–120.

[200] Yi-jie Tang and Hsin-Hsi Chen. 2014. Chinese irony corpus construction and ironic structure analysis. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING'14)*. 1269–1278.

[201] Hande Taslioglu and Pinar Karagoz. 2017. Irony detection on microposts with limited set of features. In *Proceedings of the 32nd ACM Symposium on Applied Computing (SAC'17)*. ACM, 1076–1081.

[202] Julia M. Taylor and Lawrence J. Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the 26th Meeting of the Cognitive Science Society (CogSci'04)*.

[203] Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1511–1521.

[204] Joseph Tepperman, David R. Traum, and Shrikanth Narayanan. 2006. "Yeah right": Sarcasm recognition for spoken dialogue systems. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP'06)*. International Speech Communication Association (ISCA), 1838–1841.

[205] Pyae Phyo Thu and Nwe Nwe. 2017. Impact analysis of emotion in figurative language. In *Proceedings of the 16th IEEE/ACIS International Conference on Computer and Information Science (ICIS'17)*. IEEE, 209–214.

[206] Pyae Phyo Thu and Nwe Nwe. 2017. Implementation of emotional features on satire detection. In *Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD'17)*. IEEE, 149–154.

[207] Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroglu. 2018. A computational exploration of exaggeration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. ACL, 3296–3304.

[208] Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM-A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the 4th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media (ICWSM'10)*. AAAI, 162–169.

[209] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics (ACL'14) (Long Papers)*, Vol. 1. ACL, 248–258.

[210] Piyoros Tungthamthiti, Shirai Kiyoaki, and Masnizah Mohd. 2014. Recognition of sarcasms in tweets based on concept level sentiment analysis and supervised learning approaches. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC'14)*. 404–413.

[211] Akira Utsumi. 1996. A unified theory of irony and its computational formalization. In *Proceedings of the 16th Conference on Computational Linguistics-Volume 2 (COLING'96)*. 962–967.

[212] Tony Veale and Yanfen Hao. 2007. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. In *Proceedings of the 22nd Conference of the Association for the Advancement of Artificial Intelligence (AAAI'07)*, Vol. 2. AAAI, 1471–1476.

[213] Thanh Vu, Dat Quoc Nguyen, Xuan-Son Vu, Dai Quoc Nguyen, Michael Catt, and Michael Trenell. 2018. NIHRIO at SemEval-2018 task 3: A Simple and Accurate Neural Network Model for Irony Detection in Twitter. arXiv preprint arXiv:1804.00520.
[214] Byron C. Wallace. 2015. Computational irony: A survey and new perspectives. *Artif. Intell. Rev.* 43, 4 (2015), 467–483.
[215] Byron C. Wallace, Do K. Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities, and sentiment. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*. ACL, 1035–1044.
[216] Po-Ya Angela Wang. 2013. #Irony or #sarcasm—A quantitative and qualitative study based on Twitter. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computing (PACLIC'13)*. 349–356.
[217] Zelin Wang, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter sarcasm detection exploiting a context-based model. In *Web Information Systems Engineering (WISE)*, Jianyong Wang, Wojciech Cellary, Dingding Wang, Hua Wang, Shu-Ching Chen, Tao Li, and Yanchun Zhang (Eds.). Springer, Cham, 77–91.
[218] Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psych. Rep.* 105, 2 (2009), 509–521.
[219] Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua* 116, 10 (May 2006), 1722–1743.
[220] Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua* 87, 1 (1992), 53–76.
[221] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*. ACL, 347–354.
[222] Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*. ACL, 110–114.
[223] Hongzhi Xu, Enrico Santus, Anna Laszlo, and Chu-Ren Huang. 2015. LLT-PolyU: Identifying sentiment intensity in ironic tweets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. ACL, 673–678.
[224] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. ACL, 2367–2376.
[225] Donghai Zhang, Wei Song, Lizhen Liu, Chao Du, and Xinlei Zhao. 2017. Investigations in automatic humor recognition. In *Proceedings of the 10th International Symposium on Computational Intelligence and Design (ISCID'17)*. IEEE, 272–275.
[226] Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*. 2449–2460.
[227] Renxian Zhang and Naishi Liu. 2014. Recognizing humor on Twitter. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM'14)*. ACM, 889–898.
[228] Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019. Irony detection via sentiment-based transfer learning. *Inform. Proc. & Manag.* 56, 5 (2019), 1633–1644.

Received March 2019; revised November 2019; accepted December 2019