Topics in Linguistics (2025), 26(1), pp. 174–193
10.17846/topling-2025-0008

Sarcastic comments on Reddit and Twitter

Chaya Liebeskind[1*], Anna Bączkowska[2]
[1]Jerusalem College of Technology, Israel
[2]University of Gdańsk, Poland

Key words

sarcasm
irony
BERT
sentiment analysis
social media

Abstract

The paper deals with automatic detection of sarcasm on social media (Reddit and Twitter). The objective of this paper is to explore various expressions used to encode sarcasm in comments tagged as sarcastic by social media users. The difference between sarcasm and irony and the descriptions of these terms are presented in the theoretical part. The datasets extracted from Reddit and Twitter exceeded 1.2 million comments. The research methods used in the study involved Machine Learning (ML) approaches and sophisticated deep learning algorithms (BERT, RoBERTa). It has been observed that: (i) a list of words (mostly pre-modifying amplifiers) can be identified which typically co-occur with sarcastic comments; (ii) the exclamation mark often accompanies sarcasm; (iii) sarcastic comments tend to have a negative sentiment analysis; (iv) RoBERTa models far outperformed the BERT model in detection accuracy of sarcastic posts.

1. Introduction

The eponymous concept of sarcasm is often used interchangeably with the term irony (e.g., Attardo et al. 2003, Giora et al. 2015). There is no denying that the two terms are tightly connected, yet how they relate to each other is far from obvious. To better understand the nature of sarcastic remarks, we shall start by briefly featuring the workings of irony (section 2), which will provide a conceptual platform for investigating sarcasm (section 3). Next, we will present the investigated sarcasm datasets and the computational techniques utilized t o comprehend the textual properties of sarcasm (section 4). Our study findings on the two sarcasm datasets will be presented, followed by a detailed discussion (section 5). Finally, we will outline our main conclusions and propose future study directions (section 6).

Because even humans with all of the necessary context frequently fail to detect sarcasm, it is no surprise that it is a difficult task for computers in the NLP field. One of the primary

* Corresponding author

Cite this paper: Liebeskind, C., & Bączkowska, A. (2025). Sarcastic comments on Reddit and Twitter. *Topics in Linguistics, 26*(1), 174–193. https://doi.org/10.17846/topling-2025-0008

motivations for detecting sarcasm stems from the fact that the nature of sarcasm poses a significant challenge for effective computing systems performing sentiment analysis. We believe that a necessary step in the process of effectively tackling the problem of sarcasm detection is to analyse its characteristics. Our assumption is that writers of sarcasm frequently depend not only on the shared context they have with their audience but also supplement their writing with expressions typical of sarcastic comments in order to express better the fact that they are sarcastic.

The objective of this paper is to explore various expressions used to express sarcasm. To do this, we used two different datasets containing sarcastic comments labelled with weak supervision. The first dataset (Khodak et al., 2018) contains 1 million comments collected from Reddit, half of them labelled as sarcastic. The second dataset (Cai et al., 2019) was gathered for the purpose of training a multimodal hierarchical fusion model and contains more than 20,000 comments collected from Twitter.

2. Irony

As sarcasm is often used interchangeably with irony, in this section, we will define both terms. Generally speaking, irony is claimed to be a figurative stylistic device that is not straightforward; it is hidden and encoded "between the lines", i.e., implicit (Bączkowska, 2023; Bączkowska et al., 2024). A requisite condition for irony is "the expression of feeling, attitude or evaluation" that is "a negative, hostile judgement" of the target,[1] which may take the form of derogatory and contemptuous remarks" (Grice 1989[1978], p. 53). Another common feature of irony agreed upon by a number of scholars is the fact that irony is "the statement of the opposite of what is meant" (Partington, 2007, p. 1548), "the reverse of its literal meaning" (Cutler, 1974, p. 118), or at least a salience-based (rather than literalness-based) dissimilarity between what is said and what is meant (Giora, 1995, p. 241), which runs counter to one's expectations and beliefs (Cutler, 1974; Grice, 1989[1978]; Giora, 1995), or clashes with the expectations stemming from context (Kapogianni, 2011, p. 56). Thus, by saying *You're very clever*, one may be ironic if *clever* is expected to be interpreted as its opposite (i.e., *stupid*). The contradiction need not be a full proposition, as even (non-propositional) primary interjections (e.g., *wow!*) can be ironic (Seto, 1998, p. 240). They are used in clausal or non-clausal exclamations (Biber et al., 1999, p. 909) and often "express emotional reaction to the speaker" (Biber et al., 1999, p. 140).

The use of emphatic words often marks irony (e.g., words like: *truly*, *sure*, *indeed*, *absolute*, etc.) that Seto (1998, p. 241–246) dubs echo-markers inasmuch as they make reference to an antecedent disproving proposition (often containing anaphora). Interestingly, providing an

[1] The target is an addressee who is the destination of an utterance. The target may be a sanctioned (ratified) addressee or an unratified person (e.g., when the speaker talks to person A but criticizes person B standing close to the speaker and person A).

ironic explanation after a cataphoric word is also possible, which Seto (1998, p. 242) calls non-echoic irony, as in the example he provides: *I like that. Bob smashes up my car and then expects me to pay for the repairs* (where *that* is cataphoric). Some emphatic words, such as *real* in *He's a* real *genius* (Seto, 1998, p. 243), may be read as either echoic, that is repeating its antecedent, or as a sheer amplifier that does not make reference to any previous statement, and thus it is just an "instantaneous reversal of meaning" (Seto, 1998, p. 244). Irony may involve premodifiers, as shown above, as well as heads of nominal phrases, as in *He's a genius*, wherein *genius* is ironic. As for premodifiers, along with amplifiers, the use of superlatives (*This is the wisest thing I heard*) is also a case in point, although Seto (1998, p. 244) classifies them under the rubric of syntactic devices (rather than lexical). Other syntactic devices, according to this author, comprise exclamations (*How clever!*) and (focus) topicalization (*A* fine friend *she turned out to be*), wherein words like *fine*, *lovely*, *charming*, *a lot*, etc. typically occur (Seto, 1998, p. 247).

3. Sarcasm

Closely related to irony is sarcasm, which is a trenchant, acerbic and venomous form of irony, where harsh and mocking comments are uttered to achieve a caustic effect, destroying the target interlocutor. It is an "overtly aggressive type of irony" (Attardo, 2000, p. 795), which is a face-threatening act. What makes sarcasm different from irony is the fact that the latter relies on greater wit and subtlety (Partington, 2006, p. 183), and that it is figurative (Hamamoto, 1998, p. 257), thus implicit, while the former can be straightforward (explicit) in expressing hostility or veiled (implicit), and need not be figurative. Sarcastic comments are not always ironic (Fowler, 1965, p. 535; Kreuz & Glucksberg, 1989, p. 374; Gibbs et al., 1995), yet sarcasm, in the sense of being highly critical (Mesing et al., 2012; Hanks, 2013), direct and obvious (Attardo, 2000, p. 795), and aggressive (Lee & Katz, 1998), may be used as a vehicle to convey an ironic comment. Thereby, an ironic remark may be relatively benign or extremely hostile; in the case of the latter, it is also sarcastic (Bączkowska et al., 2024). Sarcasm is believed to exert stronger emotions (Filik et al., 2019) and is used to "vent frustration" (Gibbs, 2000). In general, three essential features typify sarcasm: victim, aggressiveness and clarity/directness (Garmendia, 2018, p. 129). According to Kreuz (2020), a sarcastic tone of voice may result from a conflation of irony and hyperbole.

Dews and Winner (1999, p. 1580) concede that "Verbal irony is a form of nonliteral language in which the speaker conveys an attitude toward a person, situation, or object". While irony has a target (person, object, event, situation) and may or may not have a victim (the attacked addressee), sarcasm always has a target victim (Averbeck, 2013, p. 49; Tabacaru, 2019, p. 123; Kreuz, 2020, p. 47), who is criticized (Mesing et al., 2012). By saying *What a lovely day* one is ironic in expressing disapproval of the weather but not sarcastic (Kreuz, 2020, p. 47), unless one criticizes a person who made wrong predictions about the weather,

in which case it would be sarcasm. On the other hand, saying *You sure know a lot* (Kumon-Namakura et al., 1995, p. 7) to a person showing off his knowledge is not an example of irony, as none of the two essential requirements is met: it is neither based on (agreed/expected) insincerity (realized, for instance, by contrast) that governs most ironic remarks (the hearer-speaker believes that the original speaker knows a lot about the subject), nor on a clash between what the hearer-speaker sincerely thinks and the state of affairs in reality (which is central to verisimilar irony). It only complies with the condition of negative evaluation, possibly ridiculing the addressee. For Kapogianni (2011), negativity and ridicule would suffice to treat it as a sarcastic remark. For others, it would lack the pointed and acrimonious attitude (Filik et al., 2019) or would be insufficient in expressing ridicule and frustration (Gibbs, 2000), and thus could be seen as a negative, astringent utterance or sheer malice.[2] All these distinctive features notwithstanding, there is a feature that makes irony and sarcasm similar: they both display a tendency to induce humorous effects (Roberts & Kreuz, 1994; Jorgensen, 1996; Partington, 2006, p. 182; Norrick, 1993), yet sarcasm nurtures intentions of hostility, hurtfulness and giving pain more often than irony (Fowler, 1965, p. 535; Haiman, 1990; Dress et al., 2008).

Depending on the author, sarcasm must (Haiman, 1990, p. 181; Leggitt & Gibbs, 2000, p. 5; Sanders,[3] 2013, p. 120), may (Camp, 2012) or does not (Kapogianni, 2011)[4] entail a reading based on contrast. Accordingly, sarcasm can be literal (Kapogianni, 2011) or nonliteral (Haiman, 1990, p. 181; Camp, 2012, p. 625; Caucci & Kreuz, 2012, p. 1, 2015; Musolff, 2017, p. 96), whereas irony is conceived of as a non-literal (figurative) form of expression (which does not preclude, of course, a literal meaning encoded by what is said, which requires a figurative interpretation), i.e., one where the implied reading deviates from the literal reading. Interestingly, utterances meant to be non-sarcastic can be perceived as sarcastic when presented with particular punctuation (e.g., the exclamation mark), interjections and some adverbs (Caucci & Kreuz, 2012, p. 10). Finally, while irony is intentional or non-intentional (Muecke, 1973, p. 35, Gibbs, 2012), sarcasm is essentially intentionally malicious.

The two terms, irony and sarcasm, are often used interchangeably (Jorgensen, 1996; Lee & Katz, 1998; Attardo et al., 2003, p. 243; Sanders, 2013; Kapogianni, 2014, p. 635, Giora et al., 2015). However, sarcasm is the name which prevails in American English (particularly as a folk term, i.e., a non-technical one) to denote irony as well as sarcasm (Kapogianni, 2014,

[2] This is contrary to what Garmendia (2018, p. 27) claims who, wrongly, in our opinion, classified this example as irony.

[3] Sanders (2013, p. 120) stresses that while sarcasm rests on insincerity (and thus on incongruity), the speaker does not conceal this fact; on the contrary, it is made visible that he or she does not subscribe to the expressed opinion.

[4] This is valid for what Kapogianni (2011, p. 55) dubs *non-ironic sarcasm*, that is "a bitter comment that does not contain any conflict with reality".

p. 635; Garmendia, 2018, p. 128). This is a consequence of shifting the range of meaning of irony, which gradually started to be used to mean sheer misfortune (Attardo, 2013, p. 40), a case known in the literature as situational irony. The term irony is thus most often associated with situational irony, whereas sarcasm is prevalent in verbal irony, at least in American English (Kreuz, 2018[1996], p. 33). The discriminating facet of situational irony and sarcasm thus understood is that the former is non-verbal, whereas sarcasm is verbal and is overtly critical (Attardo, 2000, p. 795).

Interestingly, the understanding of sarcasm is not unanimous, even among American English users. As shown by Dress et al. (2008), while in Northern parts of the US, sarcasm tends to be conceptualized essentially as anchored in humour, in the Mid-Southern parts, it is perceived more often as governed by seriousness and negativity. The other potential problem stems from the mother tongue of the authors who produce sarcastic posts, i.e., when the mother tongue is not English, and the post is written in English. It has been noticed (Creusere, 1999; Giora et al., 2000) that non-native Twitter users may encounter problems with the conceptualization of sarcasm, which may partially account for erroneous post labelling.

Despite many attempts to pin down the distinctive features of irony and sarcasm, they still have fuzzy borderlines, even though over two decades have passed since Attardo (2000) made this claim. Typologically, sarcasm is most often treated as a subtype of irony (e.g., Nunberg, 2001; Gibbs, 2000; Alba-Juez & Attardo, 2014, p. 100), and this is the stand supported here, yet it is also seen as a superordinate term which encompasses irony (e.g., Camp 2012); alternatively, the two terms are viewed as too distinct notions (Caucci & Kreuz 2012, p. 1; Garmendia, 2020), or potentially distinct (Fowler, 1965, p. 535) as one concept does not automatically entail the other, or are collapsed into one category (Attardo et al., 2003; Kruger et al., 2005). Terminological issues aside, given the type of data analysed here, for this study the fine-grained distinction between them seems unnecessary, and thus they will be used as synonymous, and the theory of sarcasm will draw on the theory of irony.

4. Materials and methods

The data analysed in this study were automatically detected and gleaned from Reddit, and Twitter accounts based on tags #sarcasm, #sarcastic, #ironic or #irony. Therefore, they were tagged as such by the authors of the posts, who, as we can rather confidently surmise, are not knowledgeable in the theoretical, linguistic aspects of the two terms at hand. Consequently, the collection of comments marked as sarcastic or ironic garnered in this way may contain contexts going beyond irony/sarcasm in the linguistic sense, yet being perceived as such by social media users, that is, in the eyes of laypersons. A word of caution is thus in order with regard to the imprecise, and possibly even erroneous to some degree, make-up of any social media data extracted automatically relying on user-based labels (hashtags only). Popular parlance is not always tantamount to scholarly definitions and typologies. User-based self-

tagging and expert-based annotations may be far from unanimous or even differ substantially. As proved by Sykora et al. (2020), less than 15% of automatically detected tweets labelled with the hashtag *sarcastic* or *sarcasm* by Twitter users were tagged as such manually by an expert annotator (a linguist), and ca. 28% of contexts with the hashtag *irony* or *ironic* were consistent with expert tagging. The discrepancies only show that laypersons' and expert understandings of irony and sarcasm considerably diverge, and thus hashtags with *irony* and *sarcasm* as a clue in automatic detection are not always good indicators of the occurrence of these concepts in social media content. This limitation regardless, the undeniable advantage of automatic detection is the fact that it enables obtaining a profound insight that draws on big data (millions of contexts), which would otherwise be impossible to be investigated.

The majority of earlier work on the Natural Language Processing (NLP) task of sarcasm detection used datasets with weak supervision. The term "weak supervision" relates to the concept in which text data are only categorized as sarcastic if they satisfy a predetermined set of requirements that are assumed before data collection and analysis. This involves employing tags (such as #sarcasm and #irony) to conduct the aforementioned categorization (Ptáček et al., 2014; Khodak et al., 2018). Nonetheless, such an approach may result in noisy labelling for a variety of reasons, as shown by Oprea and Magdy (2020) and Sykora et al. (2020). Some studies have relied on parallel labelling, with human annotators providing sarcasm labels (e.g., Filatova, 2012; Abercrombie & Hovy, 2016). Yet, as Oprea and Magdy (2020) point out, such labels represent annotator perception, which may differ from author intention. A study on Twitter (González-Ibánez et al., 2011), which discovered low agreement rates between human annotators in the task of judging the sarcasm of others' tweets, supports this claim.

Another example of using Natural Language Processing tools to study a language is exploring sentiment analysis, which investigates positive or negative polarity of emotions expressed by a text. In more advanced forms of sentiment analysis, texts are not only judged based on binary emotions (positive vs negative) but on a more detailed investigation of types of emotions encoded by texts. Figurative language appears to strongly affect sentiment analysis (Ghosh et al., 2015). The results show high precision (over 90%) in predictability of negative sentiment in tweets containing irony/sarcasm.

4.1 Datasets description

In this study, we analyse and compare two weakly supervised sarcasm datasets: Reddit dataset (section 4.1.1) and Twitter dataset (4.1.2).

4.1.1 Reddit dataset

The Reddit dataset contains 1.3 million sarcastic comments and many millions more non-sarcastic comments collected by Khodak et al. (2018). The sarcastic comments were derived from Reddit comments tagged with the "/s" tag (sarcasm). This tag is commonly used by

Reddit users (henceforth "Redditors") to signal that their comment is intended to be taken in jest and should not be treated seriously.

In the Khodak et al. (2018) experiment, the Reddit weakly supervised sarcasm dataset was evaluated by estimating the false positives and false negatives percentages. A comment was deemed false positive if the "/s" tag was not actually a tag but rather a part of the sentence (possibly to mark the end of a sentence), and false negative if the author of the comment was obviously being sarcastic, at least in the opinion of the human rater. They checked two settings: the balanced and unbalanced setting and manually examined the sarcastic and non-sarcastic comments. 2.0% false negatives and 1.0% false positives were detected during this evaluation. Although the false positive rate was tolerable, the false negative rate was high compared to the sarcasm proportion (0.25%), demonstrating large heterogeneity in the working definition of sarcasm and the need for strategies to handle noisy data (i.e., data incorrectly annotated, whether false positive or false negative) in unbalanced settings.

In our paper, we utilized the balanced subset of the dataset because the balanced configuration has less noise. After removing comments containing the word sarcasm and those that were too lengthy or too short, we obtained a balanced dataset of over one million (1,010,826) Reddit comments, each provided with author, topic, and context information. Fifty percent of the comments in the balanced dataset are sarcastic (505,413).

4.1.2 Twitter dataset

The dataset of tweets that we analysed was the first dataset available for investigating multimodal sarcasm detection (Cai et al., 2019). It includes English tweets with photos and specific hashtags (such as #sarcasm) as positive examples (i.e., sarcastic) as well as English tweets with images but no such hashtags as negative examples (i.e., non-sarcastic). The following steps were taken to clean up the data: Initially, tweets with sarcasm, sarcastic, irony, and ironic as ordinary terms were removed. In addition, tweets containing words that regularly co-occur with sarcastic tweets and so may communicate sarcasm, such as jokes and humour were deleted. With a ratio of 80%:10%:10% (referring to training, development and test, respectively) the data are divided into the training set (11,174 negative examples and 8,642 positive examples), the development set (1,451 negative examples and 959 positive examples), and the test set (1,450 negative examples and 959 positive examples). In order to evaluate models more precisely, the development set and the test set were manually inspected to ensure the labels were accurate.

4.2 Computational methods used

In this work, we employed both conventional Machine Learning (ML) approaches and sophisticated deep learning algorithms to understand the text characteristics of sarcasm, as described in the next subsections.

4.2.1 Conventional Machine learning methods

To discover more sarcastic phrases and their relative relevance than can be disclosed by frequency counting, we used a standard classification process for feature extraction. We transformed a group of raw documents into a matrix of Term Frequency – Inverse Document Frequency (TF-IDF) features. TF-IDF is a commonly used statistical technique in NLP and information retrieval. It determines the importance of a word inside a document compared to a collection of documents. The TF-IDF text vectorization procedure transforms the words inside a written document into numerical representations of their significance. Then, Logistic Regression, a conventional ML classification technique, is executed. The algorithm analyses one or more continuous independent variables and one dependent variable to predict the output, category variables. We implemented our pipeline using the scikit-learn ML implementation in Python.[5] ELI5 top-level API[6] were used to provide an explanation of the linear regressor weights.

4.2.2 BERT-based masked language models

Masked language modelling (MLM) refers to the process of masking tokens in a sequence and instructing a model to fill the mask with an appropriate token. This allows the model to prioritize both the right and the left contexts. Transformer architecture serves as the foundation for BERT-based masked models.[7] A transformer is a model of deep learning that employs the process of self-attention by differently weighting the relevance of each part of the input data. Namely, BERT-based models consist of transformer encoder layers that repeatedly process the input layer by layer. Each encoder layer is responsible for generating encodings that indicate which portions of the inputs are pertinent to one another.

We used two BERT-based models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). BERT was trained on unsupervised data, as is typical for large language models, with two tasks: MLM, in which we mask one or more tokens and expect the output to be the same sentence as the sentence with the unmasked token, and Next Sentence Prediction (NSP), in which we give the model two sentences and expect the output to predict whether they follow one another or not. RoBERTa eliminates the NSP task from BERT's pre-training and incorporates dynamic masking, which causes the masked token to vary during the training epochs. In addition, the RoBERTa model is larger, was trained on more data, and employed greater batch-training sizes during the training phase.

Following pre-training, both models may be fine-tuned for particular tasks using less resources and smaller data sets. Pre-training is computationally far more expensive than fine-

[5] https://scikit-learn.org/stable/index.html

[6] https://eli5.readthedocs.io/en/latest/autodocs/eli5.html

[7] BERT stands for Bidirectional Encoder Representations from Transformers.

tuning. Consequently, in our experiments, we used pre-trained models that had already been developed, and we fine-tuned them.

4.2.3 Sentiment analysis

We used the hugging-face library's twitter-roberta-base-sentiment-latest model,[8] which was trained on 124 million tweets, to classify text sentiment into three categories: positive, negative, and neutral. The model was fine-tuned for sentiment analysis with the TweetEval benchmark (Barbieri et al., 2020).

5. Research results and discussion

5.1 An examination of the Reddit dataset

First, we compared the frequency of common words in sarcastic comments with their frequency in non-sarcastic comments. As seen on the left side of Fig. 1, the exclamation mark was the first mark that stood out. This is not surprising because, as already alluded to, Seto (1998, p. 240) observed that propositions ending with exclamation marks were not uncommon in sarcastic comments. Such propositions can be monomorphemic and syntactically independent (free-standing) interjections of exclamatory function (see Biber et al., 1999, p. 1094), such as *wow!*, that trigger strong affective states, or phrase- or sentence-long propositions ending with the exclamation mark, such as in *How clever!* (Seto, 1998, p. 247). Both types were found in our datasets. Not all posts ending with an exclamation mark in our datasets, however, are sarcastic. The examples below illustrate declarative clausal exclamatives (1–4), including two with a syntactically independent response, insert *yeah* used as discourse marker (1, 4), and a negative statement (4), all of which are ironic/sarcastic, and clausal exclamatives that are not instantiations of irony/sarcasm despite the fact that they finish with an exclamation mark, in the form of an interrogative exclamative (6) and a declarative (7).

(1) Yeah because imageboards are so much like collecting energy and resources in real life! Clausal exclamative
(2) Now you can finish your Legacy deck for cheaper!
(3) What an original post!
(4) I can't wait for 8 more years of this with Jeb!
(5) Yeah, I'm sure H2K is crying now!
(6) Who's the horrible bastard that just filmed and didn't help!
(7) You wouldn't know what it's like to be a strong disabled person pushing through a living hell every day!

[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

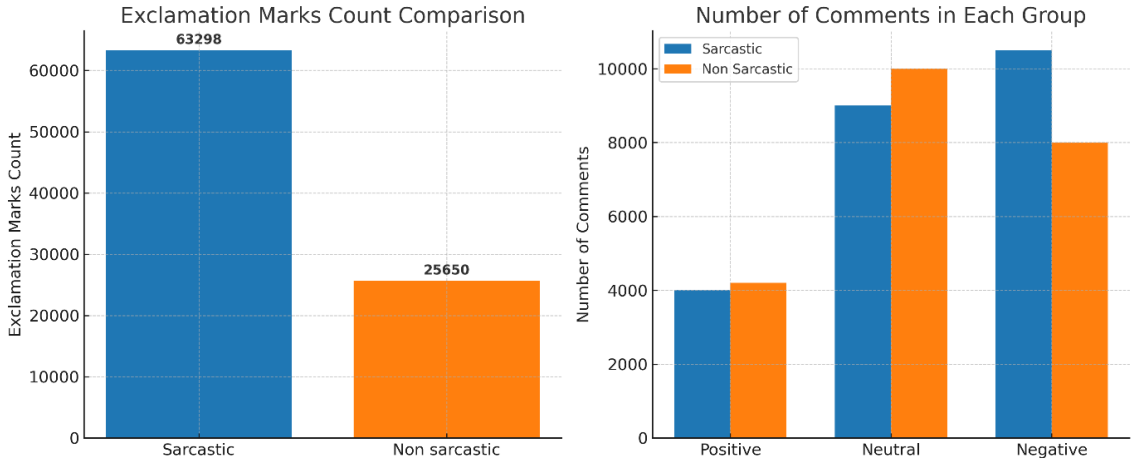


Figure 1. An evaluation of exclamation marks and sentiment of sarcastic and non-sarcastic in Reddit dataset

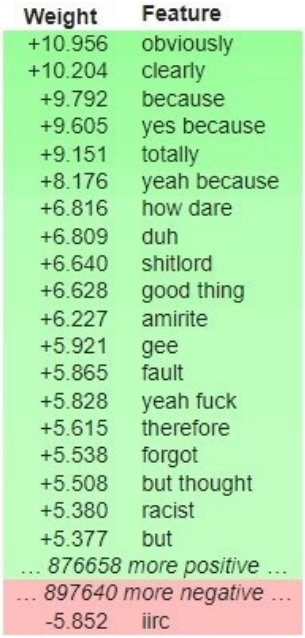
Since the exclamation mark is a token of an exclamative clause or exclamative phrase (Biber et al., 1999: 1083), one may assume that sarcastic comments are more emotional than non-sarcastic ones, as exclamations draw on emotional states. To test this theory, we compared the sentiment of sarcastic and non-sarcastic comments. Fig. 1. (on the right side) shows that sarcastic comments are less neutral and more negative. This distinction is statistically significant according to a chi-square test ($p < 0.01$).

Next, we searched for terms often used in sarcastic comments. Word clouds of sarcastic and non-sarcastic comments are shown in Fig. 2 Word clouds are graphic representations of words that emphasize terms which appear frequently (dubbed here "trigger words"). We noticed terms like *obviously*, *totally*, *right*, and *everyone knows* in the sarcastic comments. Such gradables (Biber et al., 1999, p. 555), especially those that indicate an endpoint on an imaginary scale (*totally*, *absolutely*, etc.), appear to reflect exaggeration and overconfidence in the writer's assertion, which leads us to believe that the writer is sarcastic. Linguistically, these words have the function of ironic amplifiers and are typically deployed as premodifiers in noun phrases. This observation is in line with Seto (1998, see above), who maintains that such lexemes tend to be couched in sarcastic comments which often hark back to antecedents (the so-called echo-markers).

Figure 2. Reddit dataset word clouds

To uncover more sarcastic terms and their relative importance, we then used a logistic regression model with TF-IDF (Term Frequency – Inverse Document Frequency) feature extraction for classification. The model achieved 0.81 accuracy on the training set and 0.72 accuracy on the test set, indicating relatively high performance in detecting sarcastic texts using this classification approach.

One of the key motivations for employing traditional machine learning models, such as logistic regression, lies in their interpretability: the ability to analyse and understand model outputs and, in particular, to identify the individual features (i.e., words) that most strongly influence classification outcomes. This level of transparency stands in contrast to deep learning approaches, which are often considered “black boxes” due to their complex, non-linear internal representations that hinder interpretability. Consequently, in the context of these interpretable models, we report not only performance metrics but also focus on analysing the most influential lexical features by examining the model’s learned weights (see Fig. 3 for weight values). The values in the green area correspond to features classified as positive (i.e., sarcastic), while those in the red area indicate negative (i.e., non-sarcastic) features.



Weight	Feature
+10.956	obviously
+10.204	clearly
+9.792	because
+9.605	yes because
+9.151	totally
+8.176	yeah because
+6.816	how dare
+6.809	duh
+6.640	shitlord
+6.628	good thing
+6.227	amirite
+5.921	gee
+5.865	fault
+5.828	yeah fuck
+5.615	therefore
+5.538	forgot
+5.508	but thought
+5.380	racist
+5.377	but
… 876658 more positive …	
… 897640 more negative …	
-5.852	iirc

Figure 3. Weight values of words

In the light of this, we compiled the list below by comparing the frequency of sarcastic and non-sarcastic terms in the dataset: *obviously, clearly, totally, duh, everyone knows, right because, yeah obviously, yes, because, yeah because, definitely, of course, surely, how dare, duh, gee, for sure, good, thing that, how i love, what a surprise*. As can be seen, they comprise various types of units: single word propositions expressed by interjections (*duh*); secondary interjections (*gee*); lexical words (most of them); and even phrase- (*how i love*) or sentence-based (*what a surprise*) propositions. Some of the items, such as *good thing that*, seem syntactically and/or semantically incomplete; they represent lexical bundles,[9] which, being based on frequency of co-occurrence, do not make separate and discrete units but rather latch onto neighbouring words.

Terms that appeared to be exclusive to this Reddit dataset (such as *amirite* and *shitlord*) as well as words that appear frequently (such as *because*, *fault* and *but*) were excluded. We have expanded some words in our list, for example we added the word *that* to *good things* to make a lexical bundle *good things that*. The reason for this addition is the fact that this lexical cluster was five times more frequent in sarcastic comments than the nominal phrase *good things*, which was three times more frequent in sarcastic comments. To assess the accuracy of this

[9] Lexical bundles (Biber, 1996, chapter 6) are sometimes also dubbed, *inter alia*, *n-grams* (Fletcher, 2006, p. 35), *word clusters* (Scott & Tribble, 2006) or *recurrent strings* (e.g., Altenberg, 1991). They are sequences of words that tend to co-occur, are non-idiomatic and often do not constitute an independent syntactic structure; in fact, they tend to be on the border of two different lexico-grammatical structures. They are gleaned from corpora on the basis of their frequency rather than semantic overlap (as in *torrential rain* wherein there is mutual predictability of the constitutive elements).

list, we counted how many comments containing those terms were sarcastic. We received 0.85 precision. 27,049 comments containing words that were labelled as sarcastic, which is 0.06 of the sarcastic comments.

As the next step, we looked at how these terms affected the accuracy of sarcasm detection algorithms. Since our analysis in this case suggested the presence of indicative lexical cues, we proceeded to evaluate deep learning models that incorporate contextual embeddings and are capable of capturing subtle lexical distinctions. For these models, we provide a detailed account of their classification performance across different input representation settings, highlighting the potential advantages of context-aware architectures in handling nuanced linguistic phenomena such as sarcasm. We fine-tuned two alternative models: the BERT model and the RoBERTa BERT-based model. We fine-tuned these two models using three distinct data sets. The first set consists of comments containing at least one of the terms from the previous list. They were equally divided into two categories: sarcastic and non-sarcastic. The second set consists of comments that do not include any of the words on the list, and is split in the same way as the first set. Unfiltered sarcastic and non-sarcastic comments were included in the third set.

Figure 4 illustrates a comparison between the accuracy results of the algorithms tuned on the three sets. The results of the algorithms on the first set outperform the results on the other sets, showing the importance of these terms to the model's performance and also how difficult it is to identify sarcasm without them. The model's accuracy (75%) when fine-tuned exclusively on comments containing these terms may indicate how effectively the model can learn regardless of whether the words are sarcastic or not. In all cases, the RoBERTa models outperformed the simple BERT model. When fine-tuning these models, we used the hyperparameters specified in the Pan et al. (2020), which fine-tuned a BERT model on a similar task of detecting sarcasm on Twitter.

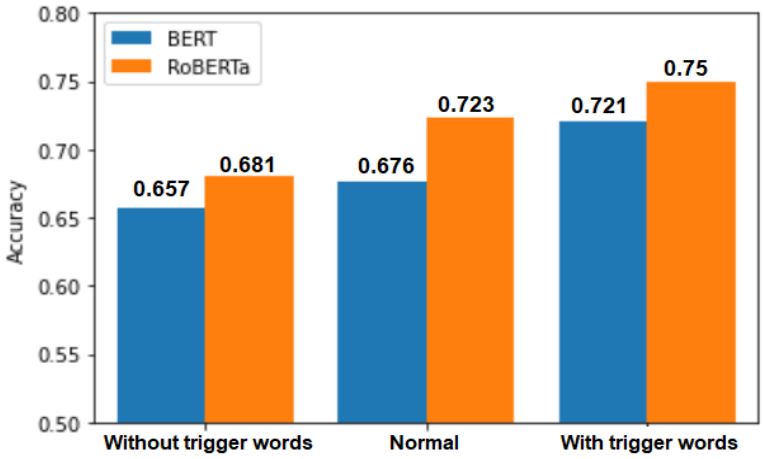


Figure 4. A comparison between the accuracy results of the algorithms tuned on the three sets

To assess the significance of the observed accuracy differences, we conducted two-proportion z-tests. For the BERT model, the improvement from the *Without trigger words* to the *Normal* configuration was not statistically significant, whereas the improvement from *Normal* to *With trigger words* was statistically significant ($p < 0.01$). In contrast, for RoBERTa, each successive configuration yielded a statistically significant improvement over the previous one ($p < 0.01$). When comparing RoBERTa and BERT across the same configurations, RoBERTa outperformed BERT with statistical significance: in the *Without trigger words* condition with $p < 0.05$, and in both the *Normal* and *With trigger words* conditions with $p < 0.01$.

5.2 An examination of the Twitter dataset

After analysing the Reddit dataset, we repeated the methodology with the Twitter dataset. As before, we first evaluated the sentiment of sarcastic and non-sarcastic tweets. On the left side of Fig. 5, we can see that the sarcastic tweets include slightly more exclamation marks, but not as many as in the Reddit database. The sentiment (see the right side of Fig. 5.), however, appears to be less neutral and more negative in sarcastic tweets, which is consistent with the findings in the Reddit dataset. This distinction is also statistically significant, as confirmed by a chi-square test with $p < 0.01$.

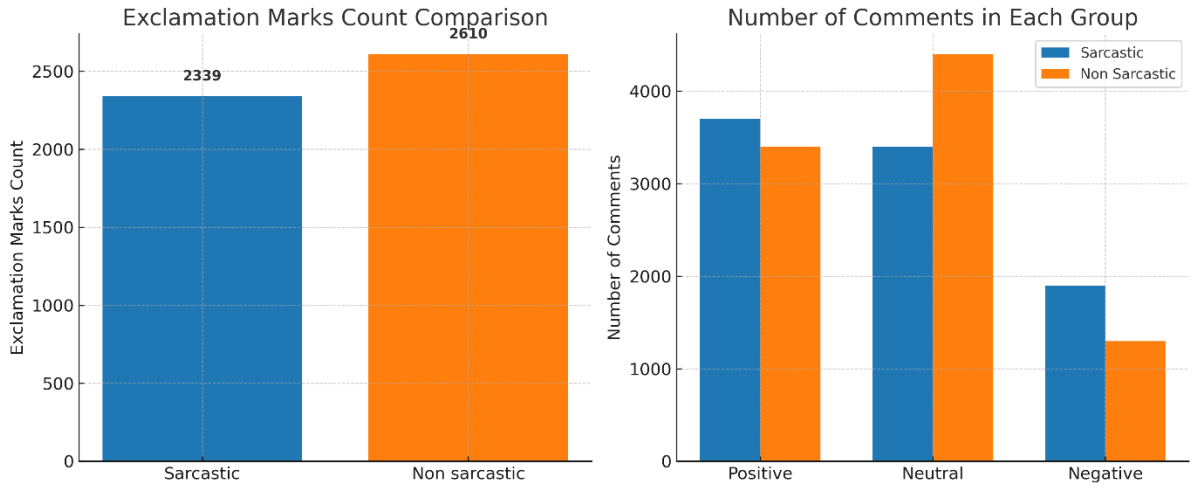


Figure 5. An evaluation of exclamation marks and the sentiment of sarcastic and non-sarcastic tweets

Following our methodology, we next generated word clouds (the left side of Fig. 6) and examined the weights of a trained logistic regression model (the right side of Fig. 6). The model achieved 0.748 accuracy on the test set.

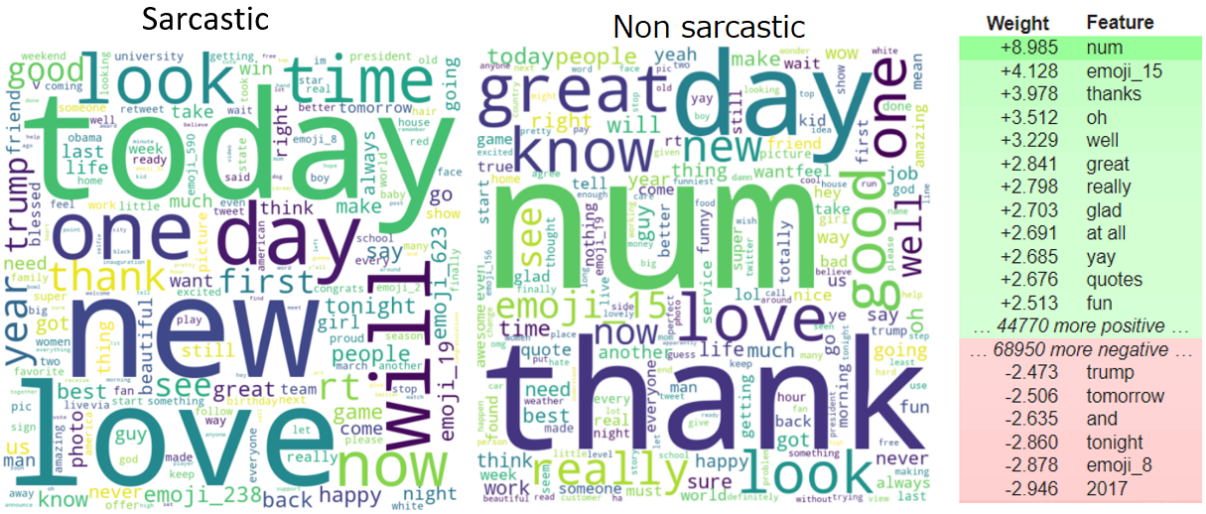


Figure 6. Word clouds of sarcastic and non-sarcastic tweets and weighed words from the logistic regression classification model

It is difficult to identify any meaningful difference between the word clouds or any words that appear to show sarcasm in the model's weights. As Oprea et al. (2020) pointed out, weakly supervised data from Twitter may be too noisy to be used for analysis and training, which our study seems to confirm.

6. Conclusion and future work

In this paper, we utilized both machine learning and neural network tools to comprehend the textual properties of sarcasm on two social media platforms (Reddit and Twitter). We have shown that a conventional logistic regression model with TF-IDF feature extraction attained a relatively high accuracy of 0.72 and 0.748 on the Reddit and Twitter test sets, respectively. We investigated which words were crucial for categorization and how these words influenced the accuracy of the sarcasm detection algorithm. Two alternative models were fine-tuned, BERT and RoBERTa, to detect sarcastic remarks based on the recognized trigger words. Of the two models, RoBERTa outperformed BERT. The results have demonstrated that resorting to trigger words, that is words which tend to pattern with sarcastic comments, such as premodifiers (amplifiers in particular), nouns as well as interjections and even syntactically/semantically incomplete units (lexical bundles), can yield very accurate results in automatic detection of sarcasm on social media. Interestingly, the exclamation mark is also a very good indicator of sarcastic comments, which was evident, particularly in data retrieved from Reddit, to a lesser extent in tweets from Twitter. Finally, sentiment analysis has demonstrated that negative polarity is typical of sarcastic remarks and that non-sarcastic remarks prevail in contexts with neutral sentiment (yet sarcastic ones are also present there).

Acknowledgment

We are grateful to Shoham Yechezkely for his significant assistance with implementing the code and running the experiments. His contribution played an important role in the technical execution of the study.

References

Abercrombie, G., & Hovy, D. (2016, August). Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop* (pp. 107–113).

Alba-Juez, L., & Attardo, S. (2014). The evaluative palette of verbal irony. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in Context* (pp. 93–115). John Benjamins.

Altenberg, B. (1991). The London-Lund Corpus: Research and applications. In *Using Corpora: Proceedings of the Seventh Annual Conference of the University of Waterloo Centre for the New OED and Text Research* (pp. 71–83).

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, *32,* 793–826.

Attardo, S. (2013). Intentionality and irony. In L. Ruiz Gurillo & B. Alvarado, B. (Eds.), *Irony and humor: From pragmatics to discourse* (pp. 39–57). John Benjamins.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, *16*(2), 243–260. https://doi.org/10.1515/humr.2003.012

Averbeck, J. M. (2013). Comparisons of Ironic and Sarcastic Arguments in Terms of Appropriateness and Effectiveness in Personal Relationships. *Argumentation and Advocacy*, *50*(1), 47–57. https://doi.org/10.1080/00028533.2013.11821809

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. T. (2020). Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics*.

Bączkowska, A. (2023). Implicit offensiveness from linguistic and computational perspectives: A study of irony and sarcasm. *Lodz Papers in Pragmatics*, *19*(2), 353–383. https://doi.org/10.1515/lpp-2023-0018

Bączkowska, A., Lewandowska-Tomaszczyk, B., Žitnik, S., Liebeskind, C., Trojszczak, M., & Valunaite Oleskeviciene, G. (2024). Implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, *20*(2), 463–483. https://doi.org/10.1515/lpp-2024-0049

Biber, D. (1996), *University language: A corpus-based study of spoken and written register*. John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Cai, Y., Cai, H., & Wan, X. (2019). Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2506–2515).

Camp, E. (2012). Sarcasm, pretense, and the semantics/pragmatics distinction. *Noûs, 46,* 587–634.

Caucci, G., & Kreuz, R. (2012). Social and paralinguistic cues to sarcasm. *Humor: International Journal of Humor Research*, *25*(1), 1–22.

Creusere, M. (1999). Theories of adults' understanding and use of irony and sarcasm: Applications to and evidence from research with children. *Developmental Review*, *19*(2), 213–262.

Cutler, A. (1974). On saying what you mean without meaning what you say. *Chicago Linguistic Society 10*, 117–27.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social functions of irony. *Discourse Processes*, *19*, 347–367.

Dews, S., & Winner, E. (1995). Muting the meaning: A social function of irony. Metaphor and Symbolic Activity, 10, 3–19.

Dews, S., & Winner, E. (1999). Obligatory processing of literal and nonliteral meanings in verbal irony. *Journal of Pragmatics 31*, 1579–1599.

Dress, M. L., Kreuz, R. J., Link, K. E., & Caucci, G. M. (2008). Regional Variation in the Use of Sarcasm. *Journal of Language and Social Psychology*, *27*(1), 71–85. https://doi.org/10.1177/0261927X07309512

Filatova, E. (2012). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. *LREC* (pp. 392–398).

Filik, R., Turcan, A., Ralph-Nearman, C., & Pitiot, A. (2019). What is the difference between irony and sarcasm? An fMRI study. *Cortex*, 115, 112–122.

Fowler, H. W. (1965). Fowler's modern English usage. 2nd ed.. Oxford University Press.

Garmendia, J. (2018). *Irony*. Cambridge University Press.

Gehweiler, E. (2010). Interjections and expletives. In A. H. Jucker & I. Taavitsainen (Eds). *Historical Pragmatics* (pp. 315–350). Mouton de Gruyter.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015). SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the*

9th International Workshop on Semantic Evaluation (SemEval 2015). https://doi.org/10.18653/v1/S15-2080

Gibbs R. (2000). Irony in talk among friends. *Metaphor and Symbol*, *15*(1–2), 5–27.

Gibbs, R. (2012). Are ironic acts deliberate? *Journal of Pragmatics*, *44*(1), 104–115.

Gibbs, R., O'Brien, J., & Doolittle, S. (1995). Inferring meanings that are not intended: speakers' intentions and irony comprehension. *Discourse Processes*, *20*(2), 187–203.

Giora, R. (1995). On irony and negation. *Discourse Processes*, *19*(2), 239–264.

Giora, R., Drucker, A., Fein, O., & Mendelson, I. (2015). Default sarcastic interpretations: On the priority of nonsalient interpretations. *Discourse Processes*, *52*(3), 173–200.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 581–586).

Grice, H. P. (1989) [1978]. Further notes on logic and conversation. In P. Grice, *Studies in the way of words* (pp. 41–57). Harvard University Press.

Haiman, J. (1990). Sarcasm as theater. *Cognitive Linguistics 1–2*, 181–205.

Hamamoto, H. (1998). Irony from a cognitive perspective. In R. Carston & S. Uchida (Eds.). *Relevance Theory: Applications and Implications* (pp. 257–270). John Benjamins.

Hanks, P. (2013). Creatively exploiting linguistic norms. In T. Veale, K. Feyaerts, & C. Forceville (Eds.). *Creativity and the agile mind. A multidisciplinary study of a multi-faceted phenomenon* (pp. 119–138). Walter de Gruyter.

Jorgensen, J. (1996). The functions of sarcastic irony in speech. *Journal of Pragmatics*, *26*(5), 613–634.

Kapogianni, E. (2011). Irony via 'surrealism'. In M. Dynel (Ed.), *The pragmatics of humour across discourse domains* (pp. 51–68). John Benjamins.

Kapogianni, E. (2014). Types and definitions of irony. In P. Stalmaszczyk (Ed.). *The Cambridge Handbook of the Philosophy of Language*. Cambridge University Press.

Devlin, J., Chang M-W, Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Khodak, M., Saunshi, N., & Vodrahalli, K. (2018). A Large Self-Annotated Corpus for Sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kreuz, R. (1996). The use of verbal irony: Cues and constraints. In J. S. Mio & A. N. Katz (Eds.), *Metaphor: Implications and Applications* (pp. 23–38). Lawrence Erlbaum.

Kreuz, R. (2018). The use of verbal irony: Cues and constraints. *Metaphor: Implications and Applications* (pp. 23–38). https://doi.org/10.4324/9781315789316-2

Kreuz, R. (2020). *Irony and Sarcasm*. MIT Press.

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: the echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General 118*, 374–386

Kruger, J., Epley, N., Parker, J., & Ng, Z. W. (2005). Egocentrism over e-mail: can we communicate as well as we think?. *Journal of personality and social psychology*, *89*(6), 925–936. https://doi.org/10.1037/0022-3514.89.6.925

Kumon-Nakamura, S., Glucksberg, S., & Brown, M. (1995). How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General, 124*, 3–21.

Lee, C. J., & Katz, A. N. (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol, 13*, 1–15.

Leggitt, J., & Gibbs, R. Jr. (2000). Emotional reactions to verbal irony. *Discourse Processes 29*, 1–24.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

Mesing, J., Williams, D., & Blasko, D. (2012). Sarcasm in relationships: hurtful or humorous? *International Journal of Psychology*, *47*, pp. 698–743.

Muecke, Douglas. 1973. The communication of verbal irony. *Journal of Literary Semantics*, *2*, 35–42.

Musolff, A. (2017). Metaphor, irony and sarcasm in public discourse. *Journal of Pragmatics 109*, 96–104.

Norrick, N. (1993). *Conversational joking: Humor in everyday talk*. Indiana University Press.

Nunberg, G. (2001). *The Way we Talk Now: Commentries on Language and Culture*. Houghton Mifflin.

Oprea, S. V., & Magdy, W. (2020). The effect of sociocultural variables on sarcasm communication online. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–22.

Pan, H., Lin, Z., Fu, P., Qi, Y., & Wang, W. (2020). Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1383–1392).

Partington, A. (2006). The linguistics of laughter. A corpus-assisted study of laughter-talk. Routledge.

Partington, A. (2007). Irony and the reversal of evaluation. *Journal of Pragmatics, 39*, 1547–1569.

Pexman, P., Reggin, L., & Lee, K. (2019). Addressing the challenge of verbal irony: getting serious about sarcasm training. *Languages*, *4*(2), 23. https://doi.org/10.3390/languages4020023

Ptáček, T., Habernal, I., & Hong, J. (2014). Sarcasm detection on Czech and English twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 213–223).

Roberts, R., & Kreuz, R. (1994). Why do people use figurative language? *Psychological Science, 5*, 159–163.

Sanders, R. (2013). The duality of speaker meaning: What makes self-repair, insincerity, and sarcasm possible. *Journal of Pragmatics 48*(1), 112–122.

Scott, M., & Tribble, C. (1997), *Textual Patterns: Keywords and Corpus Analysis in Language Education*. John Benjamins.

Seto, K-I. (1998). On non-echoic irony. In R. Carston & S. Uchida (Eds.), Relevance theory: Applications and implications (pp. 240–255). John Benjamins.

Sykora, M., Elayan, S., & Jackson, T. (2020). A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. *Big Data and Society*, *7*(2), 1–15.

Tabacaru, S. (2019). *A Multimodal Study of Sarcasm in Interactional Humor*. De Gruyter Mouton. https://doi.org/10.1515/9783110629446