Sarcasm Detection: A Comparative Study

Hamed Yaghoobian **Hamid R. Arabnia** **Khaled Rasheed**
Department of Computer Science
University of Georgia
Athens, GA, 30602, USA
{hy,hra,khaled}@uga.edu

Abstract

Sarcasm detection is the task of identifying irony[1] containing utterances in sentiment-bearing text. However, the figurative and creative nature of sarcasm poses a great challenge for affective computing systems performing sentiment analysis. This article compiles and reviews the salient work in the literature of automatic sarcasm detection. Thus far, three main paradigm shifts have occurred in the way researchers have approached this task: 1) semi-supervised pattern extraction to identify implicit sentiment, 2) use of hashtag-based supervision, and 3) incorporation of context beyond target text. In this article, we provide a comprehensive review of the datasets, approaches, trends, and issues in sarcasm and irony detection.

1 Introduction

Sarcasm poses a major challenge for sentiment analysis models (Liu et al., 2010), mainly because sarcasm enables one speaker or writer to conceal their true intention of contempt and negativity under a guise of overt positive representation. Thus, recognizing sarcasm and verbal irony is critical for understanding people's actual sentiments and beliefs (Maynard and Greenwood, 2014). The figurativeness and subtlety inherent in its sentiment display, a positive surface with a contemptuous intent (*e.g.*, *"He has the best taste in music!"*), or a negative surface with an admiring tone (*e.g.*, *"She always makes dry jokes!"*), makes the task of its identification a challenge for both humans and machines.

[1]Irony is considered an umbrella term that also covers sarcasm; distinguishing between these two rhetoric devices is a further challenge for figurative language processing (Farías et al., 2016). In short, sarcasm often bears an element of scorn and derision that irony does not (Lee and Katz, 1998).

Evidently, sarcasm and irony are well-studied phenomena in linguistics, psychology, and cognitive science. In this article, we do not survey the several representations and taxonomies of sarcasm in linguistics (Campbell and Katz, 2012; Camp, 2012; Ivanko and Pexman, 2003; Eisterhold et al., 2006; Wilson, 2006), and focus on a descriptive account of the computational attempts at automatic sarcasm detection. Empirical studies of this linguistic device refer to methods to predict if a given user-generated text is sarcastic or not. From a computational perspective, this task is formulated as a *binary classification* problem. Previous research on automated sarcasm detection has primarily focused on lexical, pragmatic resources (Kreuz and Caucci, 2007) along with interjections, punctuation, sentimental shifts, etc., found in sentences. Nonetheless, sarcasm is often manifested implicitly with no expressed lexical cues. Its identification is reliant on common sense and connotative knowledge that come naturally to most humans but makes machines struggle when extra-textual information is essentially required. Sarcastic utterances are often expressed in such nuanced ways that should be distinguished from a similar phenomenon called *humble-bragging*, which is a self-representational verbal strategy that appears as a complaint concealed within a bragging (Wittels, 2012), as in *"I am a perfectionist at times, it is so hard to deal with"*. To the best of our knowledge, there have been few computational studies that distinguish sarcasm from humble-bragging.

The remainder of this article is organized as follows. We split the literature along two discernible foci, content- and context-based methods discussed in Sections 2 and 3 respectively, and then classify empirical approaches to sarcasm detection within each section into rule-based, statistical, and deep learning-based.

2 Content-based methods

Models investigated in this section base their identification of sarcasm on lexical and pragmatic indicators in English[2] language use on social media. There is a myriad of novel and intuitive attempts in the literature that fall in this category. We review and categorize studies in this section based on approaches 2.1 (rule-based, semi-supervised and unsupervised), and features 2.2 (n-gram, sentiment, pragmatics, and patterns) used.

2.1 Rule-based

Rule-based attempts look for evidence and indicators of sarcasm and rely on those in forms of rules. Veale and Hao (2010) look for sarcastic similes (*e.g.*, "as private as a park-bench") in the specific query pattern of *"as * as a *"* on Google and using a nine-step approach reveal that 18% of unique similes are ironical.

Hashtags (or their equivalent, given the social media platform) have been utilized by users to denote sarcasm on Twitter (*e.g.*, #sarcasm, #not) or on Reddit (*e.g.*, /s). Or similarly, if the sentiment of a hashtag does not comply with the rest of the sentence, it is labeled as sarcastic.

Bharti et al. (2015) use a combination of two approaches in their study of sarcasm. They propose a parsing algorithm that looks for sentiment-bearing situations and identifies sarcasm in forms of a contradiction of negative (or positive) sentiment and positive (or negative) situation. They also look for the co-occurrence of interjection hyperbolic words like *"wow"*, *"yay"*, etc. at the start of tweets, and intensifiers like *"absolutely", "huge"* *e.g.*, "Wow, that's a huge discount, I'm not buying anything!! #sarcasm." Similarly, Riloff et al. (2013) find a positive/*negative* contrast between a sentiment and a situation helpful, and indicative of sarcasm, *e.g.*, "I'm so pleased mom *woke me up* with vacuuming my room this morning. :)". Likewise, Van Hee et al. (2018b) speculate that sentiment incongruity within an utterance signifies sarcasm. To this end, they gather all real-world concepts that carry an implicit sentiment and label them with either a "positive" or "negative" sentiment label. For example, "going to the dentist" is often associated with a negative sentiment. Although their model does not surpass the baseline, they highlight the difficulty and importance of incorporating sarcasm detection into sentiment classifiers. They view their efforts as an extension of the seminal work by Greene and Resnik (2009) to use a concept called *syntactic packaging* to demonstrate the influence of syntactic choices on the perceived implicit sentiment of news headlines.

One of the earliest work is Tepperman et al.'s that identifies sarcasm in spoken dialogues and relies heavily on cues like laughter, pauses, speaker's gender, and spectral features; their data is restricted to sarcastic utterances that contain the expression 'yeah-right'. Carvalho et al. (2009) improve the accuracy of their sarcasm model by using oral or gestural clues in user comments, such as emoticons, onomatopoeic expressions (*e.g.*, *achoo, haha, grr, ahem*) for laughter, heavy punctuation marks, quotation marks, and positive interjections. Davidov et al. (2010); Tsur et al. (2010) utilize syntactic and pattern-based linguistic features to construct their feature vectors. Barbieri et al. (2014) take a similar approach and extend previous work by relying on the inner structure of utterances such as unexpectedness, the intensity of the terms, or imbalance between registers.

2.2 Feature sets

In this section, we go over the salient textual features effectively utilized toward the detection of sarcasm. Most studies use bag-of-words to an extent. Nonetheless, in addition to these, the use of several other sets of features have been reported. Table 1 summarizes the main content-based features most commonly used in the literature. We discuss contextual features (*i.e.*, features reliant on the codification of information presented beyond text) in Section 3.

Reyes et al. (2012) introduce a set of humor-dependent or irony-dependent features related to ambiguity, unexpectedness, and emotional scenario. Ambiguity features cover structural, morphosyntactic, semantic ambiguity, while unexpectedness features gauge semantic relatedness. As we discussed, Riloff et al. (2013), in addition to a rule-based classifier, use a set of patterns, specifically positive verbs and negative situation phrases, as features. Liebrecht et al. (2013) use bigrams and trigrams and similarly, Reyes et al. (2013) look into skip-gram and character-level features. In a kindred effort, Ptáček et al. (2014) use word-

[2]Most research in sarcasm detection exists for English. Nonetheless, research in the following languages has been reported also: Utalian, Czech, Dutch, Greek, Indonesian, Chinese, and Hindi.

shape and pointedness features. Barbieri et al. (2014) include seven sets of features such as maximum/minimum/gap of intensity of adjectives and adverbs, max/min/average number of synonyms and synsets for words in the target text, and so on. Buschmeier et al. (2014) incorporate ellipsis, hyperbole, and imbalance in their set of features. Joshi et al. (2015) use features corresponding to the linguistic theory of incongruity. The features are classified into two sets: implicit and explicit incongruity-based features.

Mishra et al. (2016) propose a novel approach for investigating the salient features of sarcasm in text. They designed a set of gaze-based features such as average fixation duration, regression count, skip count, etc., based on annotations from their eye-tracking experiments. In addition, they also utilize complex gaze features based on saliency graphs, created by treating words as vertices and saccades (*i.e.*, quick jumping of gaze between two positions of rest) between a pair of words as edges.

2.3 Learning-based methods

In the following, we delve more into supervised learning, semi-supervised learning, unsupervised learning, structural and hybrid learning. A brief descriptive account of these approaches toward predictive sarcasm identification in text is given below.

2.3.1 Supervised learning

In traditional machine learning approaches, most work on statistical detection of sarcasm has relied on various combinatory forms of Random Forests (RF), Support Vector Machines (SVM), Decision trees (DT), Naïve Bayes (NB) and Neural Networks (NN) (Davidov et al., 2010; Joshi et al., 2015, 2016; Kreuz and Caucci, 2007; Reyes and Rosso, 2012; Tepperman et al., 2006; Tsur et al., 2010). For instance, González-Ibánez et al. (2011) use SVM with sequential minimal optimization (SMO) and Logistic Regression (LogR), which are usually used toward sentiment analysis, to identify discriminating features. Riloff et al. (2013) utilize a hybrid SVM system that outperformed the SVM classifier. Similarly, the use of balanced winnow algorithms to determine high-ranking features (Liebrecht et al., 2013), Naive Bayes and Decision Trees for multiple pairs of labels among irony, humor, politics, and education (Reyes et al., 2013) and fuzzy clustering for sarcasm detection (Mukherjee and Bala, 2017) are reported. Bamman and Smith (2015) present the use of binary Logistic Regression and SVM-HMM toward incorporating the sequential nature of output labels into a conversation. Likewise, Joshi et al. (2015) report that sequence labeling algorithms are more useful for conversational data as opposed to classification methods. They use SVM-HMM and SEARN as the sequence labeling algorithms. Liu et al. (2014) present a multi-strategy ensemble learning approach (MSELA) including Bagging, Boosting, etc., to handle the imbalance between sarcastic and non-sarcastic samples.

While rule-based approaches mostly rely upon lexical information and require no training, machine learning invariably makes use of training data and exploits different types of information sources (or features), such as bags of words, syntactic patterns, sentiment information or semantic relatedness. Earliest attempts in this line use similarity between word embeddings as features for sarcasm detection. Ghosh and Veale (2016) use a combination of convolutional neural networks, LSTM followed by a DNN. Van Hee et al. (2018a) propose a model that identifies sarcastic tweets and subsequently differentiates the type (out of four classes) of expressed sarcasm. The systems that were submitted for both subtasks represent a variety of neural-network-based approaches (*i.e.*, CNNs, RNNs, and (bi-)LSTMs) exploiting word and character embeddings as well as handcrafted features.

2.3.2 Semi-supervised learning

This form of machine learning, which falls between unsupervised learning and supervised learning, uses a minimal quantity of annotated (labeled) data and a large amount of un-annotated (unlabelled) data during training (Tsur et al., 2010). The presence of the unlabelled datasets and the open access to the unlabelled datasets is the feature that differentiates the semi-supervised from supervised learning. Davidov et al. (2010) employ a semi-supervised learning approach for automatic sarcasm identification using two different forms of text, tweets from Twitter, and product reviews from Amazon. A total number of 66,000 products and book reviews are collected in their study, and both syntactic and pattern-based features are extracted. The sentiment polarity of 1 to 5 is chosen on the training phase for each training data. The authors report a performance of %77 precision.

Study	Features Used
Reyes et al. (2012)	Structural, morphosyntactic and semantic ambiguity features
Tsur et al. (2010)	Internal syntactic patterns and punctuations
González-Ibánez et al. (2011)	User mentions (replies), emoticons, N-grams, dictionary- and, sentiment-lexicon-based features
Liebrecht et al. (2013)	N-grams, emotion marks, intensifiers
Hernández-Farías et al. (2015)	Length of tweet, capitalization, punctuation marks, and emoticons
Farías et al. (2016)	Lexical markers and structural features,
Mishra et al. (2016)	Cognitive features extracted from eye-movement patterns of human readers
Joshi et al. (2016)	Features based on word embedding similarity

Table 1: Features used for Statistical Classifiers

2.3.3 Unsupervised learning

Unsupervised learning in automatic sarcasm identification is still in its infancy, and most approaches are clustering-based, which are mostly applicable to pattern recognition. Nudged by the limitations and difficulties inherent in labeling the datasets (*i.e.*, time- and labor-intensivity) in supervised learning methods, researchers seek to eliminate such exertions by focusing on the development of unsupervised models. Nozza et al. (2016) propose an unsupervised framework for domain-independent irony detection. They build on probabilistic topic models originally defined for sentiment analysis. These models are extensions of the well-known Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). They propose Topic-Irony model (TIM), which is able to model irony toward different topics in a fully unsupervised setting, enabling each word in a sentence to be generated from the same irony-topic distribution. They enrich their model with a neural language lexicon derived through word embeddings. In a similar attempt, Mukherjee and Bala (2017) utilize both supervised and unsupervised settings. They use Naïve Bayes for supervised and Fuzzy C-means (FCM) clustering for unsupervised learning. Justifiably, FCM does not perform as effectively as NB.

3 Context-based models

Making sense of sarcastic expressions is heavily reliant on the background knowledge and contextual dependencies that are formally diverse. As an example, a sarcastic post from Reddit, "I'm sure Hillary would've done that, lmao." requires prior knowledge about the event, *i.e.*, familiarly with Hillary Clinton's perceived habitual behavior at the time the post was made. Similarly, sarcastic posts like "But atheism, yeah *that's* a religion!" require background knowledge, precisely due to the nature of topics like *atheism* which is often subject to extensive argumentation and is likely to provoke sarcastic construction and interpretation. The proposed models in this section utilize both content and contextual information required for sarcasm detection. In addition, there has been a growing interest in using neural language models for pre-training for various tasks in natural language processing. We go over the utilization of existing language models *e.g.*, BERT, XLNet, etc. toward sarcasm detection in section 3.1.

Wallace et al. (2014) claim that human annotators consistently rely on contextual information to make judgments regarding sarcastic intent. Accordingly, recent studies attempt to leverage various forms of contextual information mostly external to the utterance, toward more effective sarcasm identification. Intuitively, in the case of Amazon product reviews, knowing the type of books an individual typically likes might inform our judgment: someone who mostly reads and reviews Dostoevsky is statistically being ironic if they write a laudatory review of Twilight. Evidently, many people genuinely enjoy reading Twilight, and so if the review is written subtly, it will likely be difficult to discern the author's intent without this preferential background. Therefore, Mukherjee and Bala (2017) report that including features independent of the text leads to ameliorating the performance of sarcasm models. To this end, studies take three forms of context as feature: 1) author context (Hazarika et al., 2018; Bamman and Smith, 2015), 2) conversational context (Wang et al., 2015), and 3) topical context

(Ghosh and Veale, 2017). Another popular line of research utilizes various user embedding techniques that encode users' stylometric and personality features to improve their sarcasm detection models (Hazarika et al., 2018). Their model, CASCADE, utilizes user embeddings that encode stylometric and personality features of the users. When used along with content-based feature extractors such as Convolutional Neural Networks (CNNs), a significant boost in the classification performance on a large Reddit corpus is achieved. Similarly to how a user controls the degree of sarcasm in a comment, they extrapolate that the ensuing discourse of comments belonging to a particular discussion forum contains contextual information relevant to the sarcasm identification. They embed topical information that selectively incurs bias towards the degree of sarcasm present in the comments of a discussion. For example, comments on political leaders or sports matches are generally more prone to sarcasm than natural disasters. Contextual information extracted from the discourse of a discussion can also provide background knowledge or cues about the discussion topic. To extract the discourse features, they take a similar approach of document modeling performed for stylometric features.

Agrawal et al. (2020) formulate the task of sarcasm detection as a sequence classification problem by leveraging the natural shifts in various emotions over the course of a piece of text. Li et al. (2020) propose a semi-supervised method for contextual sarcasm detection in online discussion forums. They adopt author and topic sarcastic prior preference as context embedding that provides a simple yet representative background knowledge. Nimala et al. (2020) also propose an unsupervised probabilistic relational model to identify common sarcasm topics based on the sentiment distribution of words in tweets.

3.1 Sarcasm detection using pre-trained language models

Given the highlighted importance of context to capture figurative language phenomena and the difficulties of data annotation, transfer learning approaches are gaining attention in various domain adaptation problems. In particular, the utilization of pre-trained embeddings such as Global Vectors (GloVe) (Pennington et al., 2014), and ELMo (Peters et al., 2018) or leveraging Transformer seq2seq methods such as BERT (Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), etc. are witnessing a surge.

Potamias et al. (2020) propose Recurrent CNN RoBERTA (RCNN-RoBERTa), a hybrid neural architecture building on RoBERTA architecture, which is further enhanced with the employment and devise of a recurrent convolutional neural network. They report a performance with an accuracy of %79 on SARC dataset (Khodak et al., 2018). Similarly, Dadu and Pant (2020) use an ensemble of RoBERTa and ALBERT (Lan et al., 2019) on *Get it #OffMyChest* dataset (Jaidka et al., 2020) achieve a performance of %85 accuracy with $F1$ score of 0.55. Javdan et al. (2020) use BERT along with aspect-based sentiment analysis to extract the relation between context dialogue sequence and response. They obtain an F1 score of 0.73 on the Twitter dataset and 0.73 over the Reddit dataset[3]. We expect to see more studies geared toward leveraging pre-trained contextual embeddings and transformers toward sarcasm detection in the upcoming years.

Method	Acc	F1
ELMo (Peters et al., 2018)	0.70	0.70
NBSVM (Wang and Manning, 2012)	0.65	0.65
XLnet (Yang et al., 2019)	0.76	0.76
BERT-cased	0.76	0.76
RoBERTa (Liu et al., 2019)	0.77	0.77
CASCADE (Hazarika et al., 2018)	0.74	0.75
Ilic et al. (2018)	0.79	-
Khodak et al. (2018)	0.77	-
RCNN-RoBERTa (Potamias et al., 2020)	0.79	0.78

Table 2: State-of-the-art NN classifiers and results on Reddit Politics dataset

4 Datasets

This section outlines the datasets used for computational studies on sarcasm detection. Commonly, they are divided into three categories short text (*e.g.*, Tweets, Reddits), long text (*e.g.*, discussions on forums), transcripts (*e.g.*, conversational transcripts of a TV show or a call center). Short text

[3]Twitter and Reddit datasets used for in this study were provided in the shared task on Sarcasm Detection, organized at Codalab.

can contain only one (possibly sarcastic) utterance, whereas long text may contain a sarcastic sentence among other non-sarcastic sentences that could potentially function as context.

4.1 Short text

This category of data is the dominant form of expression on social media, mostly as a direct result of restriction on text length. Consequently, this type of text is rife with abbreviations to make efficient use of space on platforms such as Twitter. Two main approaches are utilized toward annotation of tweets: Manual and hashtag-based. Riloff et al. (2013); Maynard and Greenwood (2014); Mishra et al. (2016); Ptáček et al. (2014) introduce manually annotated datasets of sarcastic utterances. Most annotation approaches in the literature are conducted using hashtags to create labeled datasets. Sarcastic intent in English is commonly and culturally communicated using hashtags such as #sarcasm, #sarcastic, #not. Davidov et al. (2010); González-Ibánez et al. (2011); Reyes et al. (2012) use hashtag-based datasets of tweets. Liebrecht et al. (2013) only uses #not to collect and label their tweets. While collecting sarcastic tweets using this method is undemanding, the inclusion of non-sarcastic tweets can be challenging since tweets containing #notsarcastic may not represent a general non-sarcastic text (Bamman and Smith, 2015). Another approach is to collect the non-sarcastic tweets of users whose sarcastic tweets are also present in the dataset. To ensure collection of true sarcasm, some studies like Fersini et al. (2015) manually verified the initial hashtag-based tweets using annotators.

Reddit is the other popular platform for researchers to collect sarcasm using hashtag "/s" (Reddit's equivalent of "#sarcasm" on Twitter). Khodak et al. (2018) present SARC, a large-scale self-annotated corpus for sarcasm that contains more than a million examples of sarcastic/non-sarcastic statements made on Reddit.

4.2 Long text

Lukin and Walker (2013) use the Internet Argument Corpus (IAC) (Walker et al., 2012) which contains a set of 390,704 posts in 11,800 discussions extracted from the online debate site 4forums.com, annotated for several dialogic and argumentative markers, one of them being sarcasm. Reyes and Rosso (2014) collect a dataset of movie and book reviews, along with news articles marked with sarcasm and sentiment. In an earlier study, Reyes and Rosso (2012) garner 11,000 reviews of products with sarcastic expressions. Filatova (2012) present a corpus generation experiment where they collect regular and sarcastic Amazon product reviews. This resulting corpus can be used for identifying sarcasm on two levels: a document and a text utterance, where a text utterance can be as short as a sentence and as long as a whole document.

4.3 Transcripts and dialogues

Sarcasm is often expressed in the context of a conversation, as a response projecting contemptuous intent. Tepperman et al. (2006) uses 131 call center transcripts to look for occurrences of "yeah right" as a marker of sarcasm. Similarly, Rakov and Rosenberg (2013) through crowdsourcing collect sentences from an MTV show called "Daria." Joshi et al. (2016) also present a manually annotated transcript of the popular sitcom "Friends."

5 Conclusion

Sarcasm detection research has seen a significant surge in interest in the past few years, which justifiably calls for an investigation. This article focuses on approaches to automatic sarcasm detection in text. We discern three major paradigms in the history of sarcasm detection research: the use of hashtag-driven supervised learning toward building annotated datasets, semi-supervised pattern extraction to identify implicit sentiment, and the utilization of extra-textual information as context (*e.g.*, user's characteristic profiling). While rule-based approaches attempt to capture any indication of sarcasm in the form of rules, statistical methods use features like shifts in sentiment, specific semi-supervised patterns, etc. Deep learning techniques have also been used to incorporate context, *e.g.*, additional stylometric features of authors in conversations and the nature of discussion topics. An underlying theme of these past approaches (either in terms of rules or features) is predicated on sarcasm's contemptuous nature. Novel techniques to incorporate contextual insight have also been explored, mostly centered on the emerging direction toward utilizing language models.

References

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2020. Leveraging transitions of emotions for sarcasm detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1505–1508.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth international AAAI conference on web and social media*. Citeseer.

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*, pages 50–58.

Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2015. Parsing-based sarcasm sentiment recognition in twitter data. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1373–1380. IEEE.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49.

Elisabeth Camp. 2012. Sarcasm, pretense, and the semantics/pragmatics distinction. *Noûs*, 46(4):587–634.

John D Campbell and Albert N Katz. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480.

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 CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56.

Tanvi Dadu and Kartikey Pant. 2020. Sarcasm detection using context separators in online discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jodi Eisterhold, Salvatore Attardo, and Diana Boxer. 2006. Reactions to irony in discourse: evidence for the least disruption principle. *Journal of Pragmatics*, 38(8):1239–1256.

Delia Irazú Hernańdez Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):1–24.

Elisabetta Fersini, Federico Alberto Pozzi, and Enza Messina. 2015. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–8. IEEE.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

Irazú Hernández-Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying basic features from sentiment analysis for automatic irony detection. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 337–344. Springer.

Suzana Ilic, Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7.

Stacey L Ivanko and Penny M Pexman. 2003. Context incongruity and irony processing. *Discourse Processes*, 35(3):241–279.

Kokil Jaidka, Iknoor Singh, Jiahui Lu, Niyati Chhaya, and Lyle Ungar. 2020. A report of the cl-aff offmychest shared task: Modeling supportiveness and disclosure. In *Proceedings of the AAAI-20 Workshop on Affective Content Analysis, New York, USA, AAAI*.

Soroush Javdan, Behrouz Minaei-Bidgoli, et al. 2020. Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 67–71.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.

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 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pages 1–4.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Christopher J Lee and Albert N Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and symbol*, 13(1):1–15.

Meimei Li, Chen Lang, Min Yu, Yue Lu, Chao Liu, Jianguo Jiang, and Weiqing Huang. 2020. Scx-sd: Semi-supervised method for contextual sarcasm detection. In *International Conference on Knowledge Science, Engineering and Management*, pages 288–299. Springer.

Christine Liebrecht, Florian Kunneman, and Antal van den 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*, pages 29–37.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. 2014. Sarcasm detection in social media based on imbalanced classification. In *International Conference on Web-Age Information Management*, pages 459–471. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 Analysis in Social Media*, pages 30–40.

Diana G Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*. ELRA.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104.

Shubhadeep Mukherjee and Pradip Kumar Bala. 2017. Sarcasm detection in microblogs using naïve bayes and fuzzy clustering. *Technology in Society*, 48:19–27.

K Nimala, R Jebakumar, and M Saravanan. 2020. Sentiment topic sarcasm mixture model to distinguish sarcasm prevalent topics based on the sentiment bearing words in the tweets. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Unsupervised irony detection: a probabilistic model with word embeddings. In *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 68–76. SCITEPRESS.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, pages 1–12.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223.

Rachel Rakov and Andrew Rosenberg. 2013. " sure, i did the right thing": a system for sarcasm detection in speech. In *Interspeech*, pages 842–846.

Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision support systems*, 53(4):754–760.

Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. " yeah right": Sarcasm recognition for spoken dialogue systems. In *Ninth international conference on spoken language processing*.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*, pages 162–169. Washington, DC.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018a. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018b. We usually don't like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *ECAI*, volume 215, pages 765–770.

Marilyn Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.

Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Zelin Wang, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter sarcasm detection exploiting a context-based model. In *international conference on web information systems engineering*, pages 77–91. Springer.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.

Harris Wittels. 2012. *Humblebrag: The art of false modesty*. Grand Central Publishing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.