Automatic Sarcasm Detection: A Survey

ADITYA JOSHI, IITB-Monash Research Academy
PUSHPAK BHATTACHARYYA, Indian Institute of Technology Bombay
MARK J CARMAN, Monash University

Automatic sarcasm detection is the task of predicting sarcasm in text. This is a crucial step to sentiment analysis, considering prevalence and challenges of sarcasm in sentiment-bearing text. Beginning with an approach that used speech-based features, automatic sarcasm detection has witnessed great interest from the sentiment analysis community. This paper is a compilation of past work in automatic sarcasm detection. We observe three milestones in the research so far: semi-supervised pattern extraction to identify implicit sentiment, use of hashtag-based supervision, and incorporation of context beyond target text. In this paper, we describe datasets, approaches, trends and issues in sarcasm detection. We also discuss representative performance values, describe shared tasks and provide pointers to future work, as given in prior works. In terms of resources to understand the state-of-the-art, the survey presents several useful illustrations - most prominently, a table that summarizes past papers along different dimensions such as the types of features, annotation techniques and datasets used.

CCS Concepts: •**Information systems → Sentiment analysis;** •**Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: Sarcasm, Sentiment, Opinion, Sarcasm detection, Sentiment Analysis

ACM Reference format:
Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic Sarcasm Detection: A Survey. *ACM Comput. Surv.* 0, 0, Article 1000 (2017), 22 pages.
DOI: 00.00

1 INTRODUCTION

The Free Dictionary[1] defines sarcasm as a form of verbal irony that is intended to express contempt or ridicule. The figurative nature of sarcasm makes it an often-quoted challenge for sentiment analysis [42]. Sarcasm has a negative implied sentiment, but may not have a negative surface sentiment. A sarcastic sentence may carry positive surface sentiment (for example, '*Visiting dentists is so much fun!*'), negative surface sentiment (for example, '*His performance in Olympics has been terrible anyway*' as a response to the criticism of an Olympic medalist) or no surface sentiment (for example, the idiomatic expression '*and I am the Queen of England*'[2] is used to express sarcasm). Since sarcasm implies sentiment, detection of sarcasm in a text is crucial to predicting the correct sentiment of the text.

[1] www.thefreedictionary.com
[2] https://www.bayt.com/en/specialties/q/193911/what-s-the-meaning-or-the-best-translation-of-the-idiomatic-expression-quot-yes-and-i-am-the-queen-of-england-quot/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. 0360-0300/2017/0-ART1000 $15.00
DOI: 00.00

The challenges of sarcasm and the benefit of sarcasm detection to sentiment analysis have led to interest in automatic sarcasm detection as a research problem. Automatic sarcasm detection refers to computational approaches that predict if a given text is sarcastic. Thus, the sentence '*I love it when my son rolls his eyes at me*' should be predicted as sarcastic, while the sentence '*I love it when my son gives me a present*' should be predicted as non-sarcastic. This problem is difficult because of nuanced ways in which sarcasm may be expressed.

However, sarcasm must be distinguished from humble bragging, as in the case of the sentence '*I am having such a terrible holiday lying on the beach in the sunshine*'. Sarcasm and humble bragging are both situations where words are used to imply a sentiment that is different from their popular sentiment. Unlike humble bragging, a sarcastic sentence always has an implied negative sentiment because it intends to express contempt, as given in the definition[3]. However, to the best of our knowledge, past work does not typically distinguish between humble bragging and sarcasm, and treats humble bragging as sarcasm.

Starting with the earliest known work by [67] which deals with sarcasm detection in speech, the area has seen wide interest from the sentiment analysis community. Sarcasm detection from text has now extended to different data forms and techniques. This synergy has resulted in interesting innovations for automatic sarcasm detection. The goal of this survey paper is to provide a comprehensive summary of past work in computational sarcasm detection in order to enable new researchers to understand the state-of-the-art in this area. [71] provides a comprehensive discussion of linguistic challenges of computational irony. The paper focuses on linguistic theories and possible applications of these theories for sarcasm detection. We delve into the computational work.

The rest of the paper is organized as follows. We first describe sarcasm studies in linguistics in Section 2. To understand different aspects of the past work in automatic sarcasm detection, our paper then looks at sarcasm detection in six steps. Section 3 presents different problem definitions. Sections 4, 5 and 6 describe datasets, approaches and reported performance values, respectively. Section 7 highlights the trends underlying sarcasm detection, while Section 8 discusses recurring issues. Section 9 concludes the paper and points to future work. This survey includes tables and illustrations that serve as useful pointers to obtain a perspective on sarcasm detection research. In addition, the descriptions of shared tasks and insights for future work may be useful for a researcher in sarcasm detection and related areas.

2 SARCASM STUDIES IN LINGUISTICS

Before we begin with approaches to automatic sarcasm detection, we discuss linguistic studies pertaining to sarcasm.

Sarcasm is a form of figurative language where the literal meaning of words does not hold, and instead the opposite interpretation is intended [27]. Sarcasm is closely related to irony - in fact, it is a form of irony. [23] state that '*verbal irony is recognized by literary scholars as a technique of using incongruity to suggest a distinction between reality and expectation*'. They define two types of irony: verbal and situational. Verbal irony is irony that is expressed in words. For example, the sentence '*Your paper on grammar correction contains several grammatical errors.*' is ironic. On the other hand, situational irony is irony that arises out of a situation. For example, a situation where a scientist discovers the cure for a disease but herself succumbs to the disease before being able to apply the cure, is a situational irony.

[23] refer to sarcastic language as 'irony that is especially bitter and caustic'. There are two components of this definition: (a) presence of irony, (b) being bitter. Both together are identifying

[3]The authors thank the anonymous reviewer for suggesting the humble bragging example.

features of sarcasm. For example, '*I could not make it big in Hollywood because my writing was not bad enough*'. This example from [34] is sarcastic, because: (a) it contains an ironic statement that implies a writer in Hollywood would need to be bad at writing, (b) the appraisal in the statement is in fact bitter/contemptuous towards the entity '*Hollywood*'. Several linguistic studies describe different aspects of sarcasm:

(1) **Characteristics of sarcasm**: [11] state that sarcasm occurs along several dimensions, namely, failed expectation, pragmatic insincerity, negative tension, and the presence of a victim. [16] state that sarcasm can be understood in terms of the response it elicits. They observe that the responses to sarcasm may be laughter, no response[4], smile, sarcasm (in retort), a change of topic[5], literal reply and non-verbal reactions (a popular non-verbal reaction would be rolling one's eyes). According to [76], sarcasm arises when there is situational disparity between text and contextual information. For example, the sentence '*I love being ignored*' is understood as sarcastic due to the disparity between the contextual information that being ignored is an undesirable situation, and that the speaker claims to love it in the given sentence.

(2) **Types of sarcasm**: [10] show that there are four types of sarcasm: (i) **Propositional**: In such situations, the statement appears to be a proposition but has an implicit sentiment involved. For example '*Your plan sounds fantastic!*'. This sentence may be interpreted as non-sarcastic, if the context is not understood. (ii) **Embedded**: This type of sarcasm has an embedded incongruity in the form of words and phrases themselves. For example '*John has turned out to be such a diplomat that no one takes him seriously*'. The incongruity is embedded in the meaning of the word '*diplomat*' and rest of the sentence. (iii) **Like-prefixed**: A like-phrase provides an implied denial of the argument being made. For example, '*Like you care!*' is a common sarcastic retort. (iv) **Illocutionary**: This kind of sarcasm involves non-textual clues that indicate an attitude opposite to a sincere utterance. For example, rolling one's eyes when saying '*Yeah right*'. In such cases, prosodic variations play a role. The examples above are from [77].

(3) **Tuple-representation of sarcasm**: [29] represent sarcasm as a 6-tuple consisting of <S, H, C, u, p, p'> where: S = Speaker, H = Hearer/Listener, C = Context, u = Utterance, p = Literal Proposition, and p' = Intended Proposition.

The tuple can be read as '*Speaker S generates an utterance u in Context C meaning proposition p but intending that hearer H understands p*'. For example, if a teacher says to a student, "*That's how assignments should be done!*" and if the student knows that they have barely completed the assignment, they would understand the sarcasm. In context of the 6-tuple above, the properties of this sarcasm would be:
S: Teacher
H: Student
C: The student has not completed his/her assignment.
u: "That's how assignments should be done!"
p: The student has done a *good* job at the assignment.
p': The student has done a *bad* job at the assignment.

(4) **Echoic mention theory**: A contrary view is described in [64]. The theory states that a literal proposition may not always be intended. This can be understood with the help of the sentence '*I love it when I do not forward a chain mail and I die the next day*'. The intention

[4]Even if sarcasm is understood, it may elicit no response because it displeases or offends the listener.
[5][33] show how characters in a TV show '*Friends*' change the topic of a conversation in response to sarcasm.

of the speaker is to remind the listener of situations where chain mails do not have any result. It is through highlighting this fact that the speaker's intended ridicule of chain mails is understood. The echoic reminder theory also offers a similar perspective [39]. The echoic reminder theory states that a sarcastic statement reminds the listener of a situation they have encountered, but with clues to indicate the sarcasm. For example, the sentence '*Visits to a dentist are fun*' reminds the listener of the situation of visiting a dentist, and the popular appraisal that it is an unpleasant event.

(5) **Sarcasm as a dropped negation**: [25] states that irony/sarcasm is a form of negation in which an explicit negation marker is lacking. In other words, when one expresses sarcasm, a negation is intended, despite the lack of a negation word like '*not*'. For example, the sarcastic sentence '*Being awake at 4 am with a head-ache is fun*' is equivalent to the non-sarcastic sentence '*Being awake at 4 am with a head-ache is not fun*'. This results in the possibility that many sarcastic sentences could be converted to non-sarcastic by simply applying an appropriate negation.

(6) **Understanding sarcasm**: [24] describe how sarcasm may be understood. They state that violation of truthfulness maxims is a key for a listener to understand sarcasm. For example, '*I love being ignored*' is understood as sarcastic by a listener who believes that being ignored is not a pleasant state to be in. However, '*I love your new shirt!*' may or may not be sarcastic. The intended sarcasm cannot be understood until the listener observes that the literal meaning of the text violates truthfulness. To understand sarcasm, if any, in the sentence above, it would be essential to know information that would violate the truthfulness.

Some of these linguistic theories see a close correspondence with advances in automatic sarcasm detection. For example, the additional 'information' that is said to be necessary to violate truthfulness, is essentially the different forms of contextual information that automatic sarcasm detection aims to capture. In general, sarcasm is a verbal irony that has an intention to be mocking/ridiculing towards an entity. However, what context is required for the sarcasm to be understood forms a crucial component. Compare the sarcastic example '*I love being ignored*' with another '*I love solving math problems all day*'. The former is likely to be sarcastic for all speakers. The latter is likely to be sarcastic for most speakers. However, for authors who do really enjoy math, the statement is not sarcastic. The sarcasm in the latter may be conveyed through an author's context or paralinguistic cues (as in the case of illocutionary sarcasm). Thus, sarcasm understanding and automatic sarcasm detection are contingent on what information (or context) is known. In general, sarcasm can be summarized by a definition by [15] who state that sarcasm is "*a deliberate attempt to point out, question or ridicule attitudes and beliefs by the use of words and gestures in ways that run counter to their normal meanings*". This definition has multiple components. The first is that sarcasm is '*deliberate*' which means that it is purposefully intended by the speaker, and not an interpretation of the listener. That sarcasm '*points out, questions or ridicules*' means that there is an implied negative sentiment. For example, if a teacher catches a student cheating on a test and says, "*Your parents would be really proud today*", the implied negative sentiment is towards the listener (the student). The last part of the definition, '*in ways that run counter to their normal meanings*' highlights the relationship of sarcasm with irony. Irony is a situation in which something which was intended to have a particular result has the opposite or a very different result[6].

Thus, sarcasm consists of: (i) the use of irony, and (ii) the presence of ridicule. Based on the theories described here, understanding sarcasm (by humans or through programs) can be divided into the following components: (1) Identification of shared knowledge: The sentence '*I love being*

[6]Source: The Cambridge Dictionary

ignored' cannot be understood as sarcastic without the shared knowledge that people do not like being ignored. This specifically holds true in case of specific context. For example, the sentence '*I love solving math problems all weekend*' may be perceived as non-sarcastic by a listener who loves math or by a listener who knows that the speaker loves math. A listener, in these situations, would either look for a dropped negation or an echoic reminder, as given by theories above, (2) Identification of what constitutes ridicule: As seen in the linguistic studies, the ridicule may be conveyed through different reactions such as laughter, change of topic, etc.

Relationship with irony, deception, metaphor and humor

Sarcasm is related to other forms of incongruity or figurative language. Sarcasm has an element of ridicule that irony does not [40]. Deception also appears to be closely related to sarcasm. If a person says '*I love this soup*', they could be speaking the truth (*literal proposition*), they could be lying (*deception*) or they could be sarcastic (*sarcasm*). The difference between literal proposition and deception lies in intention of the speaker [23] while the difference between sarcasm and deception lies in shared knowledge between speaker and listener [44]. If the speaker saw a fly floating on the soup, the statement above is likely to have a sarcastic intention. Whether or not the listener understands the sarcasm depends on whether the listener saw the fly in the soup and whether the listener believes that the presence of a fly in a soup makes it bad.

[65] refer to sarcasm as a form of aggressive humor. Thus, the peculiarity that distinguishes sarcasm from another forms of incongruent expression, humor, is the element of mockery or ridicule. [23] distinguishes between metaphor and sarcasm in terms of the plausibility of the statement. They state that a metaphor is never literally plausible. For example, A says to B, '*You are an elephant*' to imply that B has a good memory is metaphorical because a human being cannot literally be an elephant. However, a sarcasm, as in the case of '*You have a very good memory*' may be plausible for people with a good memory, but sarcastic if said to a forgetful person.

These characteristics of sarcasm relate it to these linguistic expressions like humor or metaphor. It is also these characteristics such as incongruity, shared knowledge, plausibility, and ridicule that form the basis of many works in automatic sarcasm detection, as described in the forthcoming sections.

3 PROBLEM DEFINITION

We now look at how the problem of automatic sarcasm detection has been defined, in past work. The most common formulation for sarcasm detection is a **classification** task. Given a piece of text, the goal is to predict whether or not it is sarcastic. Thus, the sentence '*I love being ignored*' is to be predicted as sarcastic while the sentence '*I love being pampered*' is to be predicted as non-sarcastic. Past work varies in terms of what these output labels are. For example, trying to understand/characterize the relationship between sarcasm, irony and humor, [5] consider labels for the classifier as: politics, humor, irony and sarcasm. [59] use a similar formulation and provide pair-wise classification performance for these labels.

Other formulations for sarcasm detection have also been reported. [33] deviate from the traditional classification definition and model sarcasm detection for dialogue as a sequence labeling task. Each utterance in a dialogue is considered to be an observed unit in this sequence, whereas sarcasm labels are the hidden variables whose values need to be predicted. Thus, instead of predicting '*Yeah right*' in a conversation on its own, the system takes into account a complete conversation which includes the sentences before and after the '*yeah right*'. [22] model sarcasm detection as a **sense disambiguation** task. They state that a word may have a literal sense or a sarcastic sense. For example, the word '*amazing*' in the sentence '*Amazing to see him lose the match in 20 minutes*' is

used in a sarcastic sense, while in the sentence '*Amazing to see a brilliant film such as this*', it is used in a literal (positive) sense. Their goal is to identify the sense of a word in order to detect sarcasm. [51] model sarcasm interpretation as a monolingual machine translation task. They create a parallel corpus of sarcastic sentences and their non-sarcastic interpretations, and use two machine translation systems (based on Moses and RNN respectively) to obtain non-sarcastic interpretations of the sarcastic sentences.

Table 1 summarizes past work in automatic sarcasm detection. While several interesting observations are possible from the table, two are key: (a) tweets are the predominant text form for sarcasm detection, and (b) incorporation of extra-textual context is a recent trend in sarcasm detection.

A note on languages

Most research in sarcasm detection exists for English. However, some research in the following languages has also been reported: Chinese [43], Italian [3], Czech [53], Dutch [41], Greek [12], Indonesian [46] and Hindi [14]. Recently, [35] present a multi-lingual corpus for irony detection for French, English and Italian. The corpus uses a detailed annotation schema where ironic statements are further labeled with irony categories such as euphemism, analogy, etc.

4 DATASETS

This section describes the datasets used for experiments in sarcasm detection. We divide them into four classes: short text (typically characterized by noise and situations where length is limited by the platform, as in tweets on Twitter), long text (such as discussion forum posts), transcripts (such as transcripts of a TV show or a call center conversation), and other miscellaneous datasets. Short text can contain only one (possibly sarcastic) utterance, whereas long text may contain a sarcastic sentence among other non-sarcastic sentences. Table 2 lists past work for the three kinds of datasets.

4.1 Short text

Social media makes large-scale user-generated text accessible. However, because of restrictions on text length imposed by some of these platforms, this text tends to be short requiring authors to use abbreviations to fit their statements within the specific limit. Despite this noise, datasets of tweets have been popular for sarcasm detection because of availability of the Twitter API and popularity of Twitter as a medium. For Twitter-based datasets, two approaches to obtain annotations have been used. The first is manual annotation. [60] introduced a dataset of tweets which were manually annotated as either sarcastic or not. [47] studied sarcastic tweets and their impact to sentiment classification. They experimented with around 600 tweets which were marked for subjectivity, sentiment and sarcasm.

The second technique to create datasets is the use of hashtag-based supervision. Many approaches use hashtags in tweets as indicators of sarcasm, to create labeled datasets. The popularity of this approach (over manual annotation) can be attributed to various factors:

(1) Nobody but the author of a tweet can determine with certainty, whether the tweet was intended to be sarcastic or not. A hashtag is a label provided by authors themselves.
(2) This approach allows rapid creation of large-scale datasets since manual effort is restricted. In order to create such a dataset, tweets containing particular hashtags are labeled as sarcastic.

[13] use a dataset of tweets, which are labeled based on the presence of sarcasm-indicative hashtags such as #sarcasm, #sarcastic, #not, etc. [26] also use hashtag-based supervision for tweets. However, they retain only examples where it occurs at the end of a tweet but eliminate cases where the hashtag is a part of the running text. For example, '*#sarcasm is popular among teens*' is eliminated,

Table 1. Summary of sarcasm detection along different parameters; 'Appr.' indicates Approach, while 'Annotn.' stands for annotation technique

	Datasets			Approach			Annotation			Features					Context		
	Short Text	Long Text	Other	Rule-based	Semi-superv.	Superv	Manual	Distant	Other	Unigram	Sentiment	Pragmatic	Patterns	Other	Author	Conversation	Other
[38]			✓				✓			✓							
[13]		✓			✓		✓			✓			✓				
[68]		✓			✓		✓			✓			✓				
[69]			✓	✓					✓	✓							
[26]	✓					✓		✓		✓	✓	✓					
[19]		✓					✓										
[58]	✓					✓		✓		✓	✓	✓		✓			
[56]		✓				✓	✓			✓	✓			✓			
[41]	✓					✓		✓			✓		✓		✓		
[45]		✓			✓		✓			✓			✓				
[55]			✓			✓	✓			✓				✓			
[59]	✓					✓		✓		✓	✓	✓		✓			
[60]	✓			✓	✓		✓			✓		✓	✓				
[3]	✓					✓			✓	✓			✓				
[5]	✓					✓		✓			✓			✓			
[9]		✓				✓			✓	✓	✓	✓			✓		
[43]	✓	✓				✓		✓		✓	✓	✓		✓			
[47]	✓			✓			✓			✓	✓			✓			
[57]	✓	✓				✓		✓		✓	✓	✓		✓			
[72]		✓					✓										
[2]	✓					✓		✓		✓	✓	✓		✓	✓	✓	✓
[6]	✓				✓			✓		✓	✓		✓				
[7]	✓					✓	✓			✓		✓		✓			
[8]	✓				✓			✓		✓		✓	✓				
[18]	✓					✓		✓		✓		✓		✓			
[22]			✓				✓			✓				✓			
[20]	✓			✓	✓	✓		✓		✓		✓					
[28]	✓					✓		✓		✓	✓	✓		✓			
[32]	✓	✓				✓	✓	✓		✓	✓	✓	✓				
[36]	✓			✓			✓			✓	✓				✓		
[54]	✓					✓		✓		✓	✓			✓	✓	✓	
[73]		✓				✓	✓			✓	✓			✓		✓	✓
[75]	✓					✓			✓	✓						✓	
[1]	✓					✓	✓			✓					✓	✓	✓
[17]	✓					✓	✓	✓		✓	✓		✓	✓			
[21]	✓					✓	✓										
[33]			✓			✓	✓			✓	✓				✓	✓	
[34]			✓			✓		✓		✓				✓			
[48]	✓	✓	✓	✓			✓			✓	✓	✓		✓			
[50]	✓					✓		✓		✓	✓	✓					
[63]	✓					✓		✓							✓		

Table 2. Summary of sarcasm-labeled datasets

Text form	Related Work
Short Text	**Tweets** **Manual**: [1, 47, 48, 53, 60] **Hashtag-based**: [1–3, 5–7, 13, 18, 20, 26, 32, 36, 41, 54, 58, 59, 75]
	Reddits [37, 72, 73]
Long text	[9, 19, 43, 45, 56, 57]
Transcripts & Dialogue	[33, 55, 67]
Miscellaneous	[22, 34, 38, 48, 69]

while '*I got a parking fine on my birthday. #amazing #bestbirthdayever #sarcasm*' is included. [58] use a similar approach. [59] present a large dataset of 40000 tweets labeled as sarcastic or not, again using hashtags. [20] present hashtag-annotated dataset of tweets which they divide into three parts: 1000 trial, 4000 development and 8000 test tweets. [41] use '#not' to download and label their tweets. [5] create a dataset using hashtag-based supervision based on hashtags indicated by multiple labels namely, politics, education, sarcasm, humor and irony. Note that although these labels are not mutually exclusive by definition (for example, sarcasm is a form of irony), the presence of hashtags is used as a sole indicator of a certain label. Therefore, a tweet may be humorous and about politics, but this approach labels it purely on the basis of the hashtag present in it. For example, '*Politicians are never wrong #politics*' would get labeled as a political tweet, while '*Politicians are never wrong #sarcasm*' would get labeled as a sarcastic tweet. Other works using this approach have also been reported [1, 3, 6, 7, 32].

The Twitter API requires you to search for a keyword in a tweet (in addition to other features that it provides). Also, an author indicating, through a hashtag, that they are being sarcastic is straightforward. How to collect non-sarcastic tweets becomes tricky. Tweets containing '#notsarcastic' may be downloaded as non-sarcastic but are unlikely to have the statistical properties of general non-sarcastic text. In some cases, authors have assumed that tweets not containing the '#sarcastic' hashtag are non-sarcastic [2]. Another option is to create a list of authors of sarcastic tweets, and download their tweets not containing the sarcasm-indicative hashtag.

Hashtag-based supervision may lead to a degradation in the quality of the training data for reasons such as incorrect use of a sarcasm-indicative hashtag. To ensure quality, [18] obtain the initial label based on the presence of hashtags following which these labels are manually verified and corrected by annotators.

Twitter also provides access to additional context in terms of past tweets or author information. Hence, in order to predict sarcasm, supplementary datasets have also been used for sarcasm detection. 'Supplementary' datasets refer to text that does not need to be annotated but that will contribute to the judgment of the sarcasm detector. [36] use a supplementary set of complete Twitter timelines (limited to 3200 tweets, by Twitter) to establish context for a given dataset of tweets. [54] use a dataset of tweets, labeled by hashtag-based supervision along with a historical context of 80 tweets per author.

Like supplementary datasets, supplementary annotation (*i.e.*, annotation apart from sarcasm/non-sarcasm) has also been explored. [48] capture cognitive features based on eye-tracking. They employ annotators who are asked to determine the sentiment (and not 'sarcasm/not-sarcasm', since, as per their claim, it can result in priming) of a text. While the annotators read the text, their eye

movements are recorded by an eye-tracker. This eye-tracking information serves as supplementary annotation.

Other social media text used for sarcasm-labeled datasets includes posts from Reddit[7]. [72] create a corpus of Reddit posts of 10K sentences, from 6 Reddit topics. [73] present a dataset of Reddit comments - 5625 sentences. Similarly, [37] present a large dataset of manually labeled reddit comments, including 1.3 million sarcastic comments.

4.2 Long text

Reviews and discussion forum posts have also been used as sarcasm-labeled datasets. [45] use the Internet Argument Corpus which marks a dataset of discussion forum posts with multiple labels, one of them being sarcasm. [57] create a dataset of movie reviews, book reviews and news articles marked with sarcasm and sentiment. [56] deal with products that saw a spate of sarcastic reviews all of a sudden. Their dataset consists of 11000 reviews. [19] use a sarcasm-labeled dataset of around 1000 reviews. [9] create a labeled set of 1254 Amazon reviews, out of which 437 are ironic. [68] consider a large dataset of 66000 Amazon reviews. [43] use a dataset of reviews, comments, etc. from multiple sources such as Amazon, Twitter, Netease and Netcena. In these cases, the datasets are manually annotated because markers like hashtags are not available.

4.3 Transcripts and dialogue

Since sarcasm is a form of verbal irony and is often expressed in the context of a conversation, datasets based on transcripts and dialogue have also been reported. [67] use 131 call center transcripts. Each occurrence of 'yeah right' is marked as sarcastic or not. The goal is to identify which 'yeah right' is sarcastic. Similarly, [55] create a crowdsourced dataset of sentences from an MTV show, Daria. On similar lines, [33] report their results on a manually annotated transcript of the TV Series 'Friends'. Every 'utterance' in a scene is annotated with two labels: sarcastic or not sarcastic.

4.4 Miscellaneous datasets

In addition to the three kinds of datasets above, several other datasets have been reported. [38] use 20 sarcastic and 15 non-sarcastic book excerpts, which are marked by 101 annotators. The goal is to identify lexical indicators of sarcasm. [69] focus on identifying which similes are sarcastic. For example, the simile '*as useful as a chocolate teapot*' is to be predicted as sarcastic, while the simile '*as big as a plum*' is not[8]. Hence, they first search the web for the pattern '* as a *'. This results in 20,000 distinct similes which are then annotated. [22] use a crowdsourcing tool to obtain a non-sarcastic version of a sentence if applicable. For example '*Who doesn't love being ignored*' is expected to be corrected to '*Not many love being ignored*'. [48] create a manually labeled dataset of quotes from a website called sarcasmsociety.com. [34] create a similar dataset of quotes from GoodReads, a book recommendation website. However, in this case, they use user-determined tags to assign sarcasm labels.

5 APPROACHES

In this section, we describe approaches used for sarcasm detection. In general, approaches to sarcasm detection can be classified into: rule-based, statistical and deep learning-based approaches[9].

[7] http://www.reddit.com/

[8] Both examples are from the original paper.

[9] Obviously deep learning approaches can be considered a specific form of statistical learning, but we find it helpful notwithstanding to separate them into their own category and distinguish them from (simpler) statistical techniques such as linear classifiers.

We look at these approaches in the next sections. Following that, we describe shared tasks in research forums that deal with sarcasm detection.

5.1 Rule-based Approaches

Rule-based approaches attempt to identify sarcasm through specific evidence. This evidence is captured in the form of rules that rely on indicators of sarcasm. [69] identify sarcasm in similes using Google searches in order to determine how likely a simile is. They present a 9-step approach where at each step/rule, a simile is validated as non-sarcastic using the number of search results. To demonstrate the strength of their rules, they present an error analysis corresponding to each rule.

[47] propose that hashtag sentiment is a key indicator of sarcasm. Hashtags are often used by tweet authors to highlight sarcasm, and hence, if the sentiment expressed by a hashtag does not agree with rest of the tweet, the tweet is predicted as sarcastic. They use a hashtag tokenizer to split hashtags made of concatenated words.

[6] present two rule-based classifiers. The first uses a parse–based lexicon generation algorithm that creates parse trees of sentences and identifies situation phrases that bear sentiment. If a negative phrase occurs in a positive sentence, then the sentence is predicted as sarcastic. The second algorithm aims to capture hyperbolic sarcasm (*i.e.*, by using interjections (such as '*(*wow)' and intensifiers (such as '*absolutely*') that occur together.

[60] present rule-based classifiers that look for a positive verb and a negative situation phrase in a sentence. The set of negative situation phrases are extracted using a well-structured, iterative algorithm that begins with a bootstrapped set of positive verbs and iteratively expands both the sets (namely, positive verbs and negative situation phrases). They experiment with different configurations of rules such as restricting the order of the verb and situation phrase.

5.2 Feature Sets

In this section, we investigate the set of features that have been reported for statistical sarcasm detection. Most approaches use bag-of-words as features. However, in addition to these, several other sets of features have been reported. Table 3 summarizes them. We focus on features related to the text to be classified. Contextual features (*i.e.*, features that use information beyond the text to be classified) are described in a later section.

[68] design pattern-based features that indicate the presence of discriminative patterns (such as '*as fast as a snail*') as extracted from a large sarcasm-labeled corpus. To prevent overfitting of patterns, these pattern-based features take real values based on three situations: exact match, partial overlap and no match. [26] use sentiment lexicon-based features and pragmatic features like emoticons and user mentions. Similarly, [17] use features derived from multiple affective lexicons such as AFINN, SentiWordNet, General Inquirer, etc. In addition, they also use features based on semantic similarity, emoticons, counterfactuality, etc. [58] introduce features related to ambiguity, unexpectedness, emotional scenario, etc. Ambiguity features cover structural, morpho-syntactic, semantic ambiguity, while unexpectedness features measure semantic relatedness. [60] use a set of patterns, specifically positive verbs and negative situation phrases, as features for a classifier (in addition to a rule-based classifier). [41] introduce bigrams and trigrams as features. [59] explore skip-gram and character n-gram-based features. [3] 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, etc. [4] use similar features for irony detection. [9] incorporate ellipsis, hyperbole and imbalance in their set of features. [32] use features corresponding to the linguistic theory of incongruity. The features are classified into two sets: implicit and explicit incongruity-based features. [53] use word-shape and pointedness features. [54]

use extensions of words, number of flips, readability features in addition to contextual features. [28] present features that measure semantic relatedness between words using WordNet-based similarity. [43] introduce POS sequences and semantic imbalance as features. Since they also experiment with Chinese datasets, they use language-typical features like use of homophony, use of honorifics, etc.

Following quite a different approach, salient features of sarcastic text were investigated by [48], who 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 use complex gaze-based features based on saliency graphs which connect words in a sentence with edges representing saccades between the words.

5.3 Learning Algorithms

Most work in statistical sarcasm detection relies on different forms of Support Vector Machines (SVM) [13, 32, 38, 56, 67, 68] (or SVM-Perf as in the case of [34]). [26] use SVM and Logistic Regression, with the χ^2 test used to identify discriminating features. [60] compare rule-based techniques with a SVM-based classifier. [41] use the balanced winnow algorithm in order to determine high-ranking features. [59] use Naive Bayes and Decision Trees for multiple pairs of labels among irony, humor, politics and education. [2] use binary Logistic Regression. [75] use SVM-HMM in order to incorporate sequence nature of output labels in a conversation. Similarly, [33] validate that for conversational data, sequence labeling algorithms perform better than classification algorithms. They use SVM-HMM and SEARN as the sequence labeling algorithms. [43] compare several ensemble-based classification approaches including Bagging, Boosting, etc. and show results on five datasets. [49] use fuzzy Clustering for sarcasm detection.

5.4 Deep Learning-based Approaches

As architectures based on **deep learning** techniques gain popularity in Natural Language Programming (NLP) applications, a few such approaches have been reported for automatic sarcasm detection as well. [34] use similarity between word embeddings as features for sarcasm detection. They augment these word embedding-based features with features from four prior works. The inclusion of past features is key because they observe that using the new features alone does not suffice for good performance. [63] present a novel Convolutional Network-based architecture that learns user embeddings in addition to utterance-based embeddings. The authors state that it allows them to learn user-specific context. [21] use a combination of a Convolutional Neural Network, a Recurrent Neural Network (Long Short-Term Memory) followed by a Deep Neural Network. They compare their approach against recursive SVM, and show an improvement for the deep learning architecture. [52] investigate the use of Deep Convolutional Networks for sarcasm detection.

5.5 Shared Tasks & Benchmark Datasets

Shared tasks allow comparative evaluation of multiple approaches on a common dataset. [61] describe a shared task on sentiment analysis in Twitter, from SemEval-2014. The organizers provided multiple datasets (SMS, LiveJournal, etc.), one of which was a dataset of sarcastic tweets. It is interesting to note that the system at rank 10 gave the best performance for the dataset of sarcastic tweets (82.75, versus 77.13 which was the performance of the system that ranked highest). This best performing system was by [62]. They use features that separately capture message and contextual polarity. [20] describe a shared task from SemEval-2015 that deals with sentiment analysis of figurative language. The organizers provided a dataset of ironic and metaphorical statements labeled as positive, negative and neutral. The participants were expected to correctly identify the sentiment polarity in the case of figurative expressions like irony. The teams that participated in the

Table 3. Features used for Statistical Classifiers

	Salient Features
[68]	Sarcastic patterns, Punctuations
[26]	User mentions, emoticons, unigrams, sentiment-lexicon-based features
[58]	Ambiguity-based, semantic relatedness
[56]	N-grams, POS N-grams
[41]	N-grams, emotion marks, intensifiers
[60]	Sarcastic patterns (Positive verbs, negative phrases)
[59]	Skip-grams, Polarity skip-grams
[3]	Freq. of rarest words, max/min/avg # synsets, max/min/avg # synonyms
[5]	Synonyms, Ambiguity, Written-spoken gap
[9]	Interjection, ellipsis, hyperbole, imbalance-based
[43]	POS sequences, Semantic imbalance. Chinese-specific features such as homophones, use of honorifics
[53]	Word shape, Pointedness, etc.
[28]	Length, capitalization, semantic similarity
[32]	Unigrams, Implicit incongruity-based, Explicit incongruity-based
[54]	Readability, sentiment flips, etc.
[8]	Pattern-based features along with word-based, syntactic, punctuation-based and sentiment-related features
[17]	Affect-based features derived from multiple emotion lexicons
[34]	Features based on word embedding similarity
[48]	Cognitive features derived from eye-tracking experiments

shared task used affective resources, character n-grams, etc. The winning team used "four lexica, one that was automatically generated and three than were manually crafted. (sic)". The second shared task was a data science contest organized as a part of PAKDD 2016[10]. The competition dataset consisted of Reddit comments labeled as either sarcastic or non-sarcastic.

6 REPORTED PERFORMANCE

Table 4 presents reported values from past works. The values are not directly comparable because they are based on different datasets, experiment settings, techniques or metrics. Also, in case of some past works, only an analysis of datasets is performed without reporting metrics such as F-score, etc. These papers have not been included in the table. However, the table does provide a ballpark estimate of current performance of sarcasm detection. [26] show that unigram-based features outperform the use of a subset of words as derived from a sentiment lexicon. They compare the accuracy of the sarcasm classifier with the human ability to detect sarcasm. [56] report sentiment-based features as their top discriminating features. The Logistic Regression classifier in [55] results in an accuracy of 81.5%. [54] show that historical features along with flip-based features are the most discriminating features, and result in an accuracy of 83.46%. These are also the features presented in a rule-based setting by [36]. [17] compare their features against several reported works and show that their approach outperforms them with a F-score of 0.82.

[10]http://www.parrotanalytics.com/pacific-asia-knowledge-discovery-and-data-mining-conference-2016-contest/

Table 4. Performance Values of Sarcasm Detection; Precision/Recall/F-measures and Accuracy values are indicated in percentages

	Details	Reported Performance
[67]	Conversation transcripts	F: 70, Acc: 87
[13]	Tweets	F: 54.5 Acc: 89.6
[68]	Reviews	F: 78.8
[69]	Similes	F: 88
[26]	Tweets	A: 75.89
[58]	Irony vs general	A: 70.12, F: 65
[56]	Reviews	F: 89.1, P: 88.3, R: 89.9
[41]	Tweets	AUC: 0.76
[45]	Discussion forum posts	F: 69, P: 75, R: 62
[55]	Speech data	Acc: 81.57
[59]	Irony vs humor	F: 76
[60]	Tweets	F: 51, P: 44, R: 62
[5]	Tweets	F: 62
[9]	Reviews	F: 71.3
[47]	Tweets	F: 91.03
[2]	Tweets	Acc: 85.1
[18]	Tweets	F: 83.59, Acc: 94.17
[20]	Tweets	Cosine: 0.758, MSE: 2.117
[22]	Tweets	F: 97.5
[28]	Irony vs politics	F: 81
[32]	Tweets/Disc. Posts	F: 88.76/64
[36]	Tweets	F: 88.2
[54]	Tweets	Acc: 83.46, AUC: 0.83
[73]	Reddits	P: 0.141, F: 0.377
[75]	Tweets	Macro-F: 69.13
[1]	Tweets	AUC: 0.6
[17]	Tweets	F: 82
[33]	TV transcripts	F: 84.4
[34]	Book snippets	F: 80.47
[48]	Tweets, Quotes and Reviews	F: 75.7
[50]	Reviews	F: 75.7
[63]	Tweets	Acc: 87.2

An analysis of systems based on types of sarcasm is not available in most past work, primarily due to lack of corpora in which different kinds of sarcasm are marked. However, such an analysis is seen in [33] with respect to errors made by their system. They look at examples that their approach predicts correctly but a past approach does not. Then, they label a subset of these examples with one of the four sarcasm types. They report these values for the four types given in [10], and observe that nearly 71% require context, *i.e.*, are illocutionary or like-prefixed.

7 TRENDS IN SARCASM DETECTION

In the previous sections, we looked at the datasets, approaches and performances of past work in sarcasm detection. In this section, we detail interesting trends observed in sarcasm detection

research. These trends are represented in Figure 1. Representative work in each area are indicated in the figure, around four key milestones. Following fundamental studies, supervised/semi-supervised sarcasm classification approaches were explored. These approaches focused on using specific patterns or novel features. Then, as Twitter emerged as a viable source of data, hashtag-based supervision became popular. Recently, there is an emerging trend to use context beyond the text being classified. In the rest of this section, we describe in detail two of these trends: (a) discovery

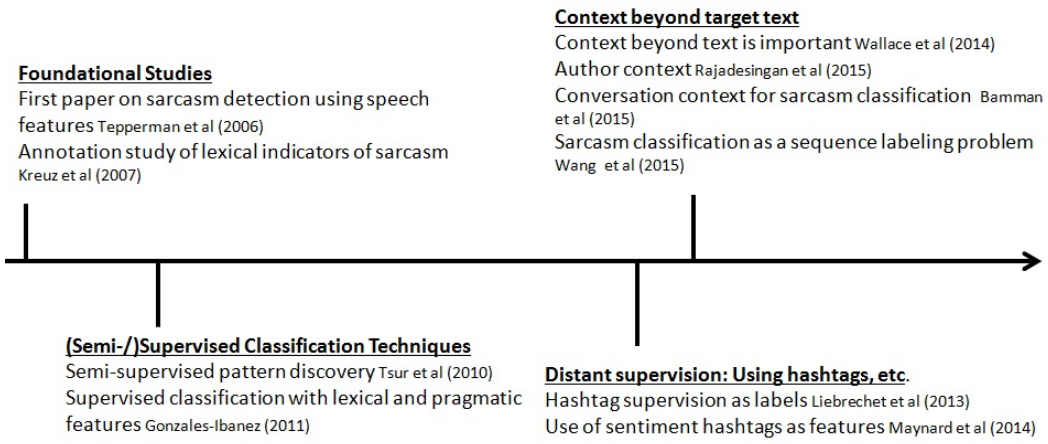

Fig. 1. Trends in Sarcasm Detection Research

of sarcastic patterns, and use of these patterns as features, and (b) use of contextual information *i.e.*, information beyond the target text for sarcasm detection.

7.1 Pattern discovery

Since sarcasm may contain implicit sentiment phrases (for example '*just got off a wonderful 12 hour flight sitting next to a crying baby*'), discovering sarcastic patterns was an early trend in sarcasm detection. Several approaches dealt with extracting patterns that are indicative of sarcasm, or carry implied sentiment. These patterns may then be used as features for a statistical classifier, or as a knowledge base for a rule-based classifier. [68] extract sarcastic patterns from a seed set of labeled sentences. They first select words that either occur more frequently than an upper threshold or less frequently than a lower threshold. They identify a large set of candidate patterns from among these extracted words. The patterns which occur discriminatively in either classes are then selected. [53] and [8] also use a similar approach for Czech and English tweets.

[60] hypothesise that sarcasm occurs due to a contrast between positive verbs and phrases indicating negative situations (as in the case of '*I love being awake at 4 am with a head-ache*' where the positive verb '*love*' is followed by the negative situation '*being awake at 4 am with a head-ache*'). To discover a lexicon of these verbs and phrases, they design an iterative algorithm. Starting with a seed set of positive verbs, they identify discriminative situation phrases that occur with these verbs in sarcastic tweets. These phrases are then used to expand the set of verbs. The algorithm iteratively appends to the lists of known verbs and phrases. The algorithm also incorporates subsumption of shorter situation phrases by longer situation phrases. For example, '*being awake at 4 am*' as in the previous example has an implicit negative sentiment in itself as compared to the longer phrase '*being awake at 4am with a head-ache*'. [32] adapt this algorithm by eliminating subsumption. [45]

begin with a seed set of nastiness and sarcasm patterns, created using Amazon Mechanical Turk. These nastiness/sarcasm patterns are statistically significant patterns (such as '*oh really*') derived from a labeled dataset. They train a high precision sarcastic post classifier, followed by a high precision non-sarcastic post classifier. These two classifiers are then used to generate a large labeled dataset, which is in turn used to train a classifier.

7.2 Role of context in sarcasm detection

An emerging trend in sarcasm detection is the use of context. The term context here refers to any information beyond the text to be predicted, and beyond common knowledge. For example, the sentence '*I love solving math problems all weekend*' may not be sarcastic to a student who loves math, but may be sarcastic to many others. This example requires context outside of the text to be classified. In the rest of this section, we describe approaches that work with examples like these. We refer to the textual unit to be classified as the 'target text'. As we will see, this context may be incorporated in a variety of ways - in general, using either supplementary data or supplementary information from the source platform providing the data. [72] is the first annotation study that highlighted the need for context for sarcasm detection. The annotators mark Reddit comments with sarcasm labels. During this annotation, annotators often request for additional context in the form of Reddit comments in the thread. The authors also present a transition matrix that shows how many times authors change their labels after this conversational context is displayed to them.

Following this observation and the promise of context for sarcasm detection, several recent approaches have looked at ways of incorporating it. The contexts that have been reported are of three types:

(1) **Author-specific context**: This type of context refers to the textual footprint of the author of the target text. For example, the statement '*Nicki Minaj.. I love her*' may be an exaggeration (hyperbolic form of sarcasm). In order to understand the sarcasm therein, information specific to the author who wrote the text is useful. [36] follow the intuition that a tweet is sarcastic either because it has words of contrasting sentiment in it, or because there is sentiment that contrasts with the author's historical sentiment'. Historical tweets by the same author are considered as the context. Named entity phrases in the target tweet are searched for in the timeline of the author in order to gather the true sentiment of the author. This historical sentiment is then used to predict whether the author is likely to be sarcastic, given the sentiment expressed towards the entity in the target tweet. [54] incorporate context about the author using the author's past tweets. This context is captured as features for a classifier. The features deal with various dimensions. They use features about the author's familiarity with Twitter (in terms of use of hashtags), familiarity with language (in terms of words and structures), and familiarity with sarcasm. [2] consider author context in features such as historical salient terms, historical topic, profile information, historical sentiment (how likely is he/she to be negative), etc. [63] capture author-specific embeddings using a neural network based architecture. The author-specific embeddings are extracted based on 1000 past tweets posted by the author.

(2) **Conversational context**: This type of context refers to text in the conversation of which the target text is a part. Consider a simple exclamation '*Yeah right, I can see that!*'. This may or may not be sarcastic. To understand the sarcasm (if any), one needs to look at the conversation that the sentence is a part of. If the sentence preceding the exclamation is '*I don't feel bad about my low grades at all*', the sarcasm in the exclamation can be inferred. [2] capture such conversational context using pair-wise Brown features between the target tweet and the previous tweet in the conversation. In addition, they also use 'audience'

features. These are features of the tweet author who responded to the target tweet. [32] show that concatenation of the previous post in a discussion forum thread along with the target post leads to an improvement in precision. [73] look at comments in the thread structure to obtain context for sarcasm detection. To do so, they use the subreddit name, and noun phrases from the thread to which the target post belongs. [75] use sequence labeling to capture conversational context. For a sequence of tweets in a conversation, they estimate the most probable sequence of three labels: happy, sad and sarcastic, for the last tweet in the sequence. A similar approach is used in [33] to predict sarcasm in every text unit in a sequence of utterances in a scene.

(3) **Topical context**: This context follows the intuition that some topics are likely to evoke sarcasm more commonly than others. For example, a tweet about a controversial topic in a political discourse is more likely to evoke sarcasm than a tweet about weather. [31] present a sarcasm topic model that uses sentiment mixture in tweets in order to discover sarcasm-prevalent topics. They note how topics like gun laws are more likely to evoke sarcasm as compared to funeral or fathers' day. [75] use topical context. To predict sarcasm in a tweet, they download tweets containing a hashtag in the tweet. Then, based on timestamps, they create a sequence of these tweets and use sequence labeling to detect sarcasm in the target tweet (the last in the sequence).

8 ISSUES IN SARCASM DETECTION

In this section, we focus on three recurring design issues that appear in different sarcasm detection works. The first deals with the quality of the **annotation**. The second issue deals with using sentiment as a **feature** for classification. Finally, the third issue lies in the context of **handling unbalanced datasets**.

8.1 Issues with Annotation

Although hashtag-based labeling can provide large-scale supervision, the quality of the dataset may be dubious. This is particularly true in the case of using #not to indicate insincere sentiment. [41] show that #not is often used to express sarcasm - while the rest of the sentence is not sufficient for identifying the sarcasm. For example, '*Looking forward to going back to school tomorrow #not*'. The speaker expresses sarcasm through #not. In most reported works that use hashtag-based supervision, the hashtag is removed in the pre-processing step. This reduces the sentence above to '*Looking forward to going back to school tomorrow*' - which may not have a sarcastic interpretation, unless the author's context is incorporated. Thus, hashtag-based supervision may cause ambiguities (or be incorrect) in some cases. To mitigate this problem, a new trend is to validate on multiple datasets - some annotated manually while others annotated through hashtags [8, 21, 32]. [21] train their deep learning-based model using a large dataset of hashtag-annotated tweets, but use a test set of manually annotated tweets.

Even in the case of manually annotated datasets, the quality of the annotation is a concern. Since sarcasm is a subjective phenomenon, the inter-annotator agreement values reported in past work are diverse. [68] indicate an agreement of 0.34. The value in the case of [18] is 0.79 while for [60], it is 0.81. [30] perform an interesting study on cross-cultural sarcasm annotation. They compare annotations by Indian and American annotators, and show that Indian annotators agree with each other more than their American counterparts. They also give examples to elicit these differences. For example, '*It's sunny outside and I am at work. Yay*' is considered sarcastic by the American annotators, but non-sarcastic by Indian annotators due to typical Indian climate.

The above observations highlight the need of framing appropriate guidelines for annotators. [70] present an annotated dataset of discussion forums with many annotations (including sarcasm). They state that for the sarcasm annotation, the annotators answer the question: '*Is the respondent using sarcasm?*'. [38] describe their annotation guidelines in terms of three questions given to the annotators: (i) how likely is this excerpt to be sarcastic, (ii) how sure are you, and (iii) why do you think it is sarcastic. Their study then analyzes lexical indicators of sarcasm from the responses to the three questions. [33] allow their annotators to look at a complete scene while annotating individual utterances. This allows them to understand the conversational context while annotating an utterance as sarcastic or not.

8.2 Issues with sentiment as a feature

Several approaches use lexical sentiment as a feature to the sarcasm classifier. It must, however, be noted that these approaches require 'surface polarity': the apparent polarity of a sentence. [6] describe a rule-based approach that predicts a sentence as sarcastic if a negative phrase occurs in a positive sentence. As described earlier, [36] use sentiment of a past tweet by the author to predict sarcasm. In a statistical classifier, surface polarity may be used directly as a feature [2, 32, 54, 58]. [59] capture polarity in terms of two emotion dimensions: activation and pleasantness. [9] use a sentiment imbalance feature that is represented by star rating of a review disagreeing with the surface polarity. [7] cascade sarcasm detection and sentiment detection, and observe an improvement of 4% in accuracy for sentiment classification, when sentiment detection is aware of sarcastic nature of text. [47] also demonstrate the impact of sarcasm detection on sentiment classification using modules from the GATE framework. Using simple rules to detect sarcasm expressed through hashtags, they improve the performance of sentiment classification by detecting sarcasm. It is important to note that [47] observe that sarcasm may not always flip the polarity. The example they quote is: '*Itfis not like I wanted to eat breakfast anyway. #sarcasm*'.

8.3 Dealing with Skewed Datasets

Sarcasm is an infrequent phenomenon of sentiment expression. This skew also reflects in datasets. [68] use a dataset with a small set of sentences are marked as sarcastic. 12.5% of tweets in the Italian dataset given by [3] are sarcastic. On the other hand, [55] present a balanced dataset of 15k tweets. In some papers, specialized techniques are used to deal with the dataset imbalances. For example, [43] present a multi-strategy ensemble learning approach. [34] use SVM-perf that performs F-score optimization. Similarly, in order to deal with sparse features and skewness of data, [73] introduce an LSS-regularization strategy. They use a sparsifying L1 regularizer over contextual features and L2-norm for bag of word features. Data imbalance also influences the choice of performance metrics reported. Since AUC is known to be a more reliable indicator of performance than F-score for skewed data, [41] report AUC for balanced as well as skewed datasets, to demonstrate the benefit of their classifier. Another methodology to ascertain benefit of a given approach withstanding data skew is by [1]. They compare performance of sarcasm classification for many datasets of different data imbalances.

9 CONCLUSION & FUTURE DIRECTIONS

Sarcasm detection research has grown significantly in the past few years, necessitating that we look back and assess the overall picture that these individual works have led to. This paper surveys approaches to automatic sarcasm detection. We observed three milestones in the history of sarcasm detection research: semi-supervised pattern extraction to identify implicit sentiment, the use of hashtag-based supervision to create large-scale annotated datasets, and the use of context beyond

target text. The paper presented illustrations containing the datasets, approaches and performance values, as reported in past work. Manually labeled datasets and datasets labeled using distant supervision are two popular techniques for creation of sarcasm-labeled datasets. We observed that rule-based approaches capture evidence of sarcasm in the form of rules such as that the sentiment of the hashtag does not match the sentiment of the rest of the tweet. Statistical approaches use features like sentiment changes, specific semi-supervised patterns, etc. Some deep learning approaches have also been reported. To incorporate context, additional features specific to the author, the conversation and the topic have been explored in the past. An underlying theme of these past approaches (either in terms of rules or features) is attempting to identify the 'irony' and 'hurtful nature' that is at the source of sarcasm.

We also highlight three issues in sarcasm detection: the quality of sarcasm annotation through manual or distant supervised datasets, the relationship between sarcasm and sentiment as a feature, and data skew in the case of sarcasm-labeled datasets. Our table that compares key past papers along dimensions such as approach, annotation approach, features, etc. should be useful to those trying to understand the current state-of-art in sarcasm detection research.

The current trend of research in sarcasm detection points to the discovery of new features and the incorporation of context. While there are manually labeled gold-standard datasets, using distant supervision to obtain labels is the prominent approach. Novel techniques to incorporate context of different forms have also been employed. Based on our survey of these works, we observe the following emerging directions:

(1) **Quality of sarcasm annotation**: Sarcasm is understood on the basis of shared knowledge. As shown in [43], sarcasm is closely related to language/culture-specific traits. Future approaches to sarcasm detection in new languages will benefit from understanding such traits, and incorporating them into their classification frameworks. [30] show that American and Indian annotators may have substantial disagreement in their sarcasm annotations. However, this sees a non-significant degradation in the performance of sarcasm detection. Since crowd-sourcing may be used for sarcasm annotation, the quality of this annotation and its impact on sarcasm classification must be evaluated on the basis of critical parameters such as cultural backgrounds.

(2) **Extraction of implicit sentiment in patterns**: Based on past work, it is well-established that sarcasm is closely linked to sentiment incongruity [32]. Several related works exist for detection of implicit sentiment in sentences, as in the case of '*The phone gets heated quickly*' v/s '*The induction cooktop gets heated quickly*'. This will help sarcasm detection, following the line of semi-supervised pattern discovery.

(3) **Analysis based on types of sarcasm**: As noted in the survey, past work does not report which of the types of sarcasm are correctly handled by existing systems. A dataset which labels sarcastic sentences into one of the four types, and then studies the performance of various systems on each of these types will be helpful. Future work can benefit from reporting which types of sarcasm are proving to be difficult for different approaches.

(4) **Sarcasm versus irony classification**: Sarcasm and irony are closely related and most work so far considers them to be the same. However, some recent work has dealt with understanding the differences between the two. [74] present findings of a data analysis to understand differences between sarcasm and irony. According to them, aggressiveness is the distinguishing factor between the two. [66] present a set of classifiers that distinguish between sarcasm and irony. They describe an analysis of structural and affective features in tweets. An important observation that they make is the peculiarity of the hashtag '#not' as a negation marker for sarcasm.

(5) **Linguistic basis for sarcasm detection**: Many sarcasm theories, except the theory of dropped negation (described in Section 2) have not been explored as means for sarcasm detection. [66] show that the hashtag '#not' plays a distinct role in sarcastic tweets. This may have correlations with this theory of dropped negation. Approaches grounded in linguistic theories may yield good results.
(6) **Coverage of different forms of sarcasm**: In Section 2, we described four species of sarcasm: propositional, lexical, like-prefixed and illocutionary sarcasm. We observe that current approaches are limited in handling the last two forms of sarcasm: like-prefixed and illocutionary. Future work may focus on these forms of sarcasm.
(7) **Extraction of Contextual Information using Deep learning-based architectures**: Very few approaches have explored deep learning-based architectures so far. As shown in [63], context embeddings can be captured. Embeddings derived from other forms of context may be useful to capture the additional shared knowledge (say, user or conversation-specific knowledge) that is required to understand certain forms of sarcasm.

ACKNOWLEDGMENTS

The authors would like to thank the Reviewers and the Editorial Committee of ACM Computing Surveys for their immensely useful feedback. This paper was written based on the understanding gained during the primary author's doctoral research on computational sarcasm. Therefore, the authors are grateful to their collaborators during the course of this research: Vaibhav Tripathi, Vinita Sharma, Diptesh Kanojia, Kevin Patel, Anoop Kunchukuttan, Samarth Agrawal, Anupam Khattri, Prayas Jain, Pranav Goel, Jaya Saraswati, Rajita Shukla and Meghna Singh. The authors also thank Abhijit Mishra and Joe Cheri Ross for their feedback on the paper.

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. *ACL 2016* (2016), 107.
[2] David Bamman and Noah A Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Ninth International AAAI Conference on Web and Social Media*.
[3] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2014. Italian irony detection in twitter: a first approach. In *The First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA*. 28–32.
[4] Francesco Barbieri and Horacio Saggion. 2014. Modelling Irony in Twitter: Feature Analysis and Evaluation. In *LREC*. 4258–4264.
[5] Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling Sarcasm in Twitter, a Novel Approach. *ACL 2014* (2014), 50.
[6] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2015. Parsing-based Sarcasm Sentiment Recognition in Twitter Data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 1373–1380.
[7] Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 1594–1597.
[8] Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Sarcasm Detection in Twitter:" All Your Products Are Incredibly Amazing!!!"-Are They Really?. In *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.
[9] Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. *ACL 2014* (2014), 42.
[10] Elisabeth Camp. 2012. Sarcasm, Pretense, and The Semantics/Pragmatics Distinction*. *Noûs* 46, 4 (2012), 587–634.
[11] John D Campbell and Albert N Katz. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes* 49, 6 (2012), 459–480.
[12] 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. *Engineering Applications of Artificial Intelligence* (2016). DOI: http://dx.doi.org/10.1016/j.engappai.2016.01.007

[13] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 107–116.

[14] Nikita Desai and Anandkumar D Dave. 2016. Sarcasm Detection in Hindi sentences using Support Vector machine. *International Journal* 4, 7 (2016).

[15] Madhav Deshpande. 2002. *Indian linguistic studies: festschrift in honor of George Cardona.* Motilal Banarsidass Publ.

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

[17] Delia Irazú Hernańdez Farías, Viviana Patti, and Paolo Rosso. 2016. Irony Detection in Twitter: The Role of Affective Content. *ACM Trans. Internet Technol.* 16, 3, Article 19 (July 2016), 24 pages.

[18] Elisabetta Fersini, Federico Alberto Pozzi, and Enza Messina. 2015. Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*. IEEE, 1–8.

[19] Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *LREC*. 392–398.

[20] Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Int. Workshop on Semantic Evaluation (SemEval-2015)*.

[21] Aniruddha Ghosh and Tony Veale. 2016. Fracking Sarcasm using Neural Network. *WASSA NAACL 2016* (2016).

[22] Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. In *EMNLP*.

[23] Raymond W Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding.* Cambridge University Press.

[24] Raymond W. Gibbs and Jennifer O'Brien. 1991. Psychological aspects of irony understanding. *Journal of Pragmatics* 16, 6 (1991), 523 – 530. DOI: http://dx.doi.org/10.1016/0378-2166(91)90101-3

[25] Rachel Giora. 1995. On irony and negation. *Discourse processes* 19, 2 (1995), 239–264.

[26] 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: short papers-Volume 2*. Association for Computational Linguistics, 581–586.

[27] H Paul Grice, Peter Cole, and Jerry L Morgan. 1975. Syntax and semantics. *Logic and conversation* 3 (1975), 41–58.

[28] Irazú Hernández-Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In *Pattern Recognition and Image Analysis*. Springer, 337–344.

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

[30] Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How Do Cultural Differences Impact the Quality of Sarcasm Annotation?: A Case Study of Indian Annotators and American Text. *LaTeCH 2016* (2016), 95.

[31] Aditya Joshi, Prayas Jain, Pushpak Bhattacharyya, and Mark James Carman. 2016. fiWho would have thought of that!fi: A Hierarchical Topic Model for Extraction of Sarcasm-prevalent Topics and Sarcasm Detection. In *Workshop on Extra-Propositional Meaning at COLING 2016*. 1.

[32] 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*, Vol. 2. 757–762.

[33] Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark Carman. 2016. Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series 'Friends'. *CoNLL 2016* (2016), 146.

[34] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are Word Embedding-based Features for Sarcasm Detection? *EMNLP 2016* (2016).

[35] Jihen Karoui, Farah Benamara, 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. European Chapter of the Association for Computational Linguistics.

[36] Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2015. Your Sentiment Precedes You: Using an authorfis historical tweets to predict sarcasm. In *6TH WORKSHOP ON COMPUTATIONAL APPROACHES TO SUBJECTIVITY, SENTIMENT AND SOCIAL MEDIA ANALYSIS WASSA 2015*. 25.

[37] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A Large Self-Annotated Corpus for Sarcasm. *arXiv preprint arXiv:1704.05579* (2017).

[38] 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*. Association for Computational Linguistics, 1–4.

[39] Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General* 118, 4 (1989), 374.

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

[41] CC Liebrecht, FA Kunneman, and APJ van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. (2013).

[42] Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 2 (2010), 627–666.

[43] 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*. Springer, 459–471.

[44] Debra L Long and Arthur C Graesser. 1988. Wit and humor in discourse processing. *Discourse processes* 11, 1 (1988), 35–60.

[45] 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*. 30–40.

[46] Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on*. IEEE, 195–198.

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

[48] 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*.

[49] Shubhadeep Mukherjee and Pradip Kumar Bala. 2017. Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. *Technology in Society* 48 (2017), 19–27.

[50] Smaranda Muresan, Roberto Gonzalez-Ibanez, Debanjan Ghosh, and Nina Wacholder. 2016. Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology* (2016).

[51] Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting Sarcasm with Sentiment Based Monolingual Machine Translation. *arXiv preprint arXiv:1704.06836* (2017).

[52] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic Tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815* (2016).

[53] Tomáš Ptácek, Ivan Habernal, and Jun Hong. 2014. Sarcasm Detection on Czech and English Twitter. In *Proceedings COLING 2014*. COLING.

[54] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on Twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 97–106.

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

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

[57] 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 (2014), 595–614.

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

[59] Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47, 1 (2013), 239–268.

[60] 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 *EMNLP*. 704–714.

[61] Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. Dublin, Ireland, 73–80.

[62] José Saias. 2014. Senti. ue: Tweet overall sentiment classification approach for SemEval-2014 task 9. Association for Computational Linguistics.

[63] Amir Silvio, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling Context with User Embeddings for Sarcasm Detection in Social Media. *CoNLL 2016* (2016), 167.

[64] Dan Sperber. 1984. Verbal irony: Pretense or echoic mention? *Journal of Experimental Psychology: General* 113, 1 (1984), 130–136.

[65] Stefan Stieger, Anton K Formann, and Christoph Burger. 2011. Humor styles and their relationship to explicit and implicit self-esteem. *Personality and Individual Differences* 50, 5 (2011), 747–750.

[66] Emilio Sulis, Delia Farías, Delia Irazú Hernańdez, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems* 108 (2016), 132 – 143. New Avenues in Knowledge Bases for Natural Language Processing.
[67] Joseph Tepperman, David R Traum, and Shrikanth Narayanan. 2006. " yeah right": sarcasm recognition for spoken dialogue systems.. In *INTERSPEECH*. Citeseer.
[68] 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*.
[69] Tony Veale and Yanfen Hao. 2010. Detecting Ironic Intent in Creative Comparisons.. In *European Conference on Artificial Intelligence*, Vol. 215. 765–770.
[70] Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *LREC*. 812–817.
[71] Byron C Wallace. 2013. Computational irony: A survey and new perspectives. *Artificial Intelligence Review* 43, 4 (2013), 467–483.
[72] Byron C Wallace, Laura Kertz Do Kook Choe, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 512–516.
[73] Do Kook Choe Wallace, Byron C and Eugene Charniak. 2015. Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. In *ACL*.
[74] Po-Ya Angela Wang. 2013. #Irony or #Sarcasmfi!? A Quantitative and Qualitative Study Based on Twitter. In *27th Pacific Asia Conference on Language, Information, and Computation*. 349–256.
[75] Zelin Wang, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter Sarcasm Detection Exploiting a Context-Based Model. In *Web Information Systems Engineering–WISE 2015*. Springer, 77–91.
[76] Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua* 116, 10 (2006), 1722–1743.
[77] Zsófia Zvolenszky. 2012. A Gricean rearrangement of epithets. *Twenty years of theoretical linguistics in Budapest* (2012), 183.

Received September 2016; revised March 2017; accepted July 2017