IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 16, 2022, accepted May 1, 2022, date of publication May 13, 2022, date of current version May 31, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3174862

Sarcasm Over Time and Across Platforms: Does the Way We Express Sarcasm Change?

MONDHER BOUAZIZI[1], (Member, IEEE), AND TOMOAKI OHTSUKI[2], (Senior Member, IEEE)
[1]Ohtsuki Laboratory, Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan
[2]Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan

Corresponding author: Mondher Bouazizi (bouazizi@ohtsuki.ics.keio.ac.jp)

ABSTRACT Sarcasm is a sophisticated form of speech used to convey a message other than the apparent one. To date, there are numerous papers that have discussed the idea of automatic sarcasm detection and how it could be used for sentiment analysis improvement. The objective of this paper is to provide non-experts with a comprehensive overview of the state of research in this field and the main findings regarding sarcasm detection. Therefore, in this paper, we survey the state-of-the-art work done in this field, we recapitulate the research effort done, with focus on the more recent works, and we present the expected performance out of the proposed works. Nevertheless, we study in detail how this form of speech is used in different platforms, and how the way we express it evolves over time. We also discuss the proposition that suggests that sarcasm is a polarity switcher for sentiment analysis. To achieve these goals, we run some experiments on 3 different data sets, collected from 3 different platforms, and compare how sarcasm is employed in each. These platforms are Twitter, Reddit, and some news websites. Our experiments show that the way sarcasm is expressed is highly dependent on language mastery and the platform used. For instance, in the Twitter data set, whose users vary widely in age, language mastery, and understanding what sarcasm means, the overall precision of detection of sarcastic statement reaches 89.31%. In the reddit data set, the precision of detection of such statements is about 55.33%, and in the news data set, the precision reaches over 96.67%. Our experiments also show that, to a great extent, it is safe to affirm that sarcasm, when employed, switches the polarity of a given piece of text: for the 3 platforms presented above, sarcasm has been a polarity switcher for 89.3%, 89.1%, and 92.0% of their respective instances.

INDEX TERMS Deep learning, machine learning, sarcasm detection, sentiment analysis.

I. INTRODUCTION

With the rapid growth of user-generated content on the Internet, companies, organizations, and research institutions and centers have been studying this type of data for several purposes. Part of this work has been interested in the interaction between the Internet users, the types of exchange of information they do, and even the nature of the relationships they build. However, most of the interest has targeted the content of the data they share, for it being the most rich in terms of information embedded. Several studies have been conducted on the content of the user-generated data. One particular type study performed on these data is referred to as sentiment analysis. Sentiment analysis refers to the process of automatically identifying the opinion embedded within a given piece of text. Roughly speaking, sentiment analysis has as a first goal the detection of the sentiment polarity of the text. By sentiment polarity, we mean identifying whether the author of the text has a positive attitude towards its subject or a negative one (or sometimes a neutral one). Sentiment analysis has several usages, varying from the identification of users' opinion on a product or service [1]–[3] to their voting intent on upcoming elections, etc. With its maturity, sentiment analysis-related research has deviated from bringing to the table novel approaches to perform the task, towards applications of this technique in cases such as the US presidential elections [4], the Coronavirus pandemic [5]–[7], and critical events [8], etc.

That being said, despite how sophisticated the approaches proposed in the literature are, sentiment analysis, after all, relies mostly on words and expressions used in a text to identify its polarity. However, appearances might be misleading. This is the case when non-straightforward and indirect forms of speeches such as sarcasm are employed.

The associate editor coordinating the review of this manuscript and approving it for publication was Joey Tianyi Zhou.

Sarcasm has had an increase in usage in social media over the last few years, with a multitude of accounts named after it spreading sarcastic statements, which are shared and re-posted by millions of users. Sarcasm has been used by normal users as well as public figures in online debates or when addressing a public event or hot and controversial topics.

The Collins online dictionary[1] defines sarcasm as a *"speech or writing which actually means the opposite of what it seems to say"*. Cambridge dictionary[2] defines it as *"the use of remarks that clearly means the opposite of what they say"*. Several previous works have shown that sarcasm is one of the most common reasons of misclassification when sentiment analysis is performed [9], [10]. With reference to the definitions mentioned above, sarcasm can roughly be defined as saying the opposite of what is meant, an idea which we discuss in more detail later on in this paper. Sarcasm is being widely used for several reasons, the most important among them being how pertinent and expressive sarcastic statements sound: As discussed by R. Giora [11], direct negation can sometimes be vague and not very expressive. It can also sound very serious and face-threatening, and can sometimes sound dull, or not conveying the feeling of the person talking. Sarcasm, and irony in general, is less serious, yet very expressive. It also conveys more than just the idea which one wants to negate. For instance, when one wants to express his being annoyed of someone else, he might use the expression *"You are so funny!"*. This expression doe not only tell the other party that he is not funny, but also gives him the impression that the person is annoyed by his stories. Likewise, Camp [12] analyzed sarcasm in terms of meaning inversion, and distinguished 4 sub-classes of sarcasm, individuated in terms of the target of inversion:

- Propositional sarcasm: which is more like the traditional model suggests where sarcasm is as simple as saying the contrary of a proposition that would have been expressed by a sincere utterance.
- Lexical sarcasm: which delivers an inverted compositional value for only a single expression or part of the sentence.
- ''Like''-prefixed sarcasm: which commits the speaker to the emphatic epistemic denial of a declarative utterance's focal content.
- Illocutionary sarcasm: which expresses an attitude which is the opposite of one that a sincere utterance would have expressed.

She also concluded that 3 of these classes raise serious challenges for a standard implicature analysis.

With that in mind, despite the common thought that a person's way of expressing himself is an idiosyncrasy, a complex and unique way for himself, it is undoubtedly more accurate to assume that the way we behave is learned from others, the way we talk is, more or less, a combination of what we have heard and expressed in the past [13]. This has been addressed by the developmental psychologists and proven to be very accurate [14]. Sarcasm, for instance, is one of the most sophisticated forms of speech that, ironically, many people are less creative when trying to employ. Some suggest that such form of speech requires high Intelligence Quotient (IQ) to be able to express, let alone to catch and understand [15]. In [16], the authors suggested that people tend to rely on cheap or lazy cues to detect it. Therefore, it has been noticeable that many so-called ''sarcastic statements'' on social media are simple iterations on already-established sarcastic statements. In other words, most of what casual Internet users create as sarcastic statements are modification of previously created ones to fit in a given context. This idea of lack of creativity is the basis of several previously proposed works on the automatic sarcasm detection on texts collected from social media and microblogging websites such as Facebook and Twitter [17]–[19]. These works rely on what they refer to as ''sarcastic patterns'' to identify such common expressions used to express sarcasm.

The use of sarcastic patterns to locate sarcastic statements has had very good results on data collected from online social networks and microblogging websites such as Twitter. However, the question yet to answer would be whether or not such idea can be used to identify sarcasm on more structured data types or on texts written by people with higher language mastery.

To recapitulate, the objective of this survey paper is to provide non-experts with a comprehensive overview of the state of research in this field and the main findings extracted by the researchers regarding sarcasm detection. Nevertheless, in this paper, we try to answer the following 3 questions:

[Q1] Does the way people express sarcasm differ from one platform to another, and does it depend on the level of ''mastery'' of the language?

[Q2] Does the way people express sarcasm evolve over time, in particular, on social media where sarcastic statements are ''driven'' by some influential users?

[Q3] Is it safe to affirm that, for a given piece of text, if sarcasm is employed, the overall polarity of that text is the opposite of the apparent one?

The remainder of this paper is structured as follows: Section II describes briefly the state of the art of existing work that dealt with tasks of sentiment analysis and sarcasm detection. In Section III, we present some of the work related to sarcasm in other fields, as well as the main findings that could hint to possible ways to understand sarcasm, thus to detect it. In Section IV, we explore in more detail the existing works on automatic sarcasm detection, covering the data sets built for this task, the methods used, the features extracted from the data to identify sarcasm and the reported results. In Section V, we summarize the main challenges and problems that are still open for research in this field. In Section VI, we present our experiment specifications including a description of the data sets we have used and the software and hardware environments. We describe our

[1]https://www.collinsdictionary.com
[2]https://dictionary.cambridge.org

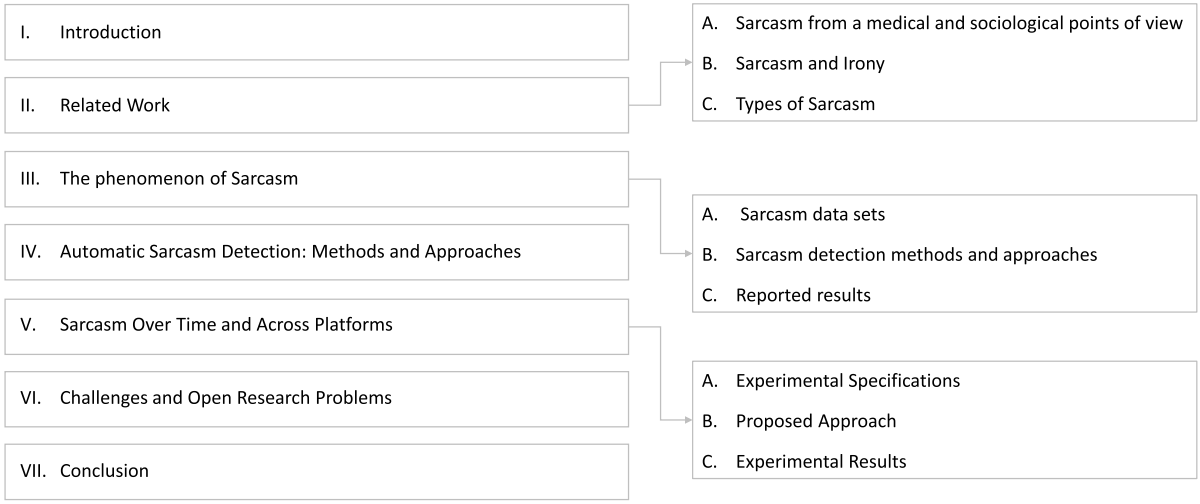


FIGURE 1. Outline of this article.

TABLE 1. List of acronyms and their corresponding full forms.

Acronym	Full form
ALS	Amyotrophic Lateral Sclerosis
API	Application Programming Interface
AUC	Area Under the Curve
AWD-LSTM	Averaged Stochastic Gradient Descent Weight-Dropped LSTM
Bi-LSTM	Bidirectional Long Short Term Memory
CNN	Convolutional Neural Network
DL	Deep Learning
HMM	Hidden Markov Models
IQ	Intelligence Quotient
kNN	k-Nearest Neighbors
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processessing
NN	Neural Network
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TPR	True Positive Rate
URL	Uniform Resource Locator
VC	Vapnik–Chervonenkis

implementations for sarcasm detection. We then address the questions **[Q1]**, **[Q2]**, and **[Q3]**, show the different experiments we have run, and discuss the results obtained.

Finally, in Section VII, we conclude this paper and propose possible directions for future work. For more readability, the outline of this article is shown in Figure 1. In addition, the most used acronyms and their full forms are shown in Table 1.

II. RELATED WORK

A. SENTIMENT ANALYSIS

As described in Section I, sarcasm detection has almost completely been associated with the idea of sentiment analysis enhancement. Sentiment analysis has a long history that goes back to the ancient Greece [20], [21]. However, this kind of analysis was very basic and non-robust and does not qualify as scientific. This is because it did not follow the scientific method which has been established centuries later. Nonetheless, it did not benefit from the currently existing technology which has allowed for massive application of sentiment analysis on real large-scale problems. From the science point of view, the first journal on public opinion mining was published in the year 1931 [22]. However, sentiment analysis as we know now has been defined by Lee who co-authored later the work [23] and who is considered to be one of the founders of the field of "Sentiment Analysis" in the early 2000s. Pioneered by the work of Pang *et al.* [23], the idea of using machine learning for sentiment analysis has been massively adopted, and the vast majority of works in the field have opted for the use of machine learning. Research on sentiment analysis has since then known an exponential growth, with many approaches revolving around the same basic idea proposed afterwards. According to [20], over 99% of scientic papers on sentiment analysis have been published after the year 2004.

The spread of social media over the last two decades has resulted in an exponential growth of user-generated data, a perfect material for application of sentiment analysis. This is because user-generated data are regarded as raw Internet users' opinions, which can be analyzed for various objectives. For instance, sentiment analysis has been used typically for collecting, analyzing, and aggregating people's opinions about products [24], [25] or movies [2], or services [26]–[28]. Nevertheless, works such as that of Akcora *et al.* [29] were proposed identify major changes in public opinion over the time, and spot the news that led to breakpoints in public opinion.

Twitter, being one of the most popular platforms for people to share their thoughts in relatively short texts, has attracted most of the attention in the last few years. Approaches such as that of Boia *et al.* [30] and that of Manuel *et al.* [31] used

non-textual features (such as emoticons and slang) to classify tweets and online texts or to attribute sentiment scores to them.

Newer trends in sentiment analysis include multi-class sentiment analysis [32]–[35], sentiment quantification [36]–[40], and applications of sentiment analysis in general such as the US presidential elections [4], the Coronavirus pandemic [5]–[7], etc.

B. SARCASM DETECTION

Sarcasm detection for sentiment analysis improvement is relatively a novel field. To the best of our knowledge, the first published work to introduce this task was that of Tepperman *et al.* [41]. However, its being applied on vocal data makes it a bit different from the rest of the works discussed here, and from the task we will undego later on. Kreuz and Caucci [42] introduced this task for written text. In their work, they used unigrams to identify sarcastic phrases and sentences present in excerpts from long narratives. Their approach, despite being naive, was a start point for several works to come in the next years.

Tsur *et al.* [43] and Davidov *et al.* [17] have introduced a semi-supervised approach to detect sarcastic statements on Twitter and Amazon. They introduced the concept of sarcastic patterns to refer to generic expressions that are commonly used in sarcastic statements. This idea has been polished further in other works such as those of Lukin and Walker [44], Liebrecht *et al.* [45], Barbieri *et al.* [46] and Bouazizi and Ohtsuki [47].

Nevertheless, other works have been introduced in the next years. Some of the works used *n*-grams [48], while other used other types of features such as sentiment features [49], [50]. More advanced ones make use of the context within which the text message was posted, that being temporal, conversational, psychological or behavioral [51], [52].

In addition, with the advances in the field of Deep Learning (DL), several approaches were proposed to detect sarcasm using this technology which has proven to outperform conventional Machine Learning (ML) in classification tasks. Poria *et al.* [53] have proposed a model to extract sentiment, emotion, and personality features for sarcasm detection.

On a related context, Twitter has been the main platform which has been studied on sarcasm detection. This is because of the reasons we have introduced in the previous section, in addition to the openness of this platform and the ease of access to its users' generated content, via its streaming Application Programming Interface (API). However, several works were introduced to detect sarcasm on other platforms and types of texts, such as Amazon reviews [43], Reddit posts and comments [54], news articles [55], etc.

Research on automatic sarcasm detection is still ongoing, and the results obtained in this field are promising, and have real-world applications to improve sentiment analysis.

In the next Section, we will address further the current state of the studies on sarcasm in different fields, before we tackle the works in the field of automatic sarcasm detection. We will describe in more detail the techniques used, the results obtained and the main findings.

III. THE PHENOMENON OF SARCASM

A. SARCASM FROM A MEDICAL AND SOCIOLOGICAL POINTS OF VIEW

Sarcasm detection, as addressed in this paper, relates to the process of using Natural Language Processing (NLP) techniques and tools to automatically detect sarcasm from social media and other online user-generated content sources. However, sarcasm has nonetheless been studied in other fields such as the medical field, in particular from a neurological perspective. For instance, damage to brain cells, and mental deficiency limit largely one's ability to capture sarcasm [56], which might lead to undesirable consequences. Sarcasm is by definition used to express criticism, quite often in a non-aggressive way. Not being able to understand it does not only reveal mental deficiency, but also leads to miscommunication and incorrect interpretations of intentions. With that in mind, people with mental health issues such as dementia or even non-demented problems [57] share common behavior regarding the processing of indirect forms of speech. Staios *et al.* [58] explored sarcasm detection in amyotrophic lateral sclerosis using ecologically valid measures. They have shown that Amyotrophic Lateral Sclerosis (ALS) patients exhibit cognitive deficits, including being unable to understand and detect sarcastic and paradoxical sarcastic statements, both being sophisticated forms of speech. This goes along with other observations that suggest that sarcasm requires high IQ to understand [15], even though low IQ does not necessarily mean having neurological problems.

Nevertheless, sarcasm has also been studied from a psychological and sociological perspectives. Sarcasm usage could imply a certain degree of closeness between the speaker and the hearer [59]. Not only does it reflect the nature of the speaker as indirect and humorous, but also it has effects on the hearer, whether this effect is positive [60] or negative [59].

Whether sarcasm is a polarity switcher has also been addressed in few works [47]. In addition, despite being confused in many works with the concept of irony, Littman and Mey [61] suggested that sarcasm and irony are not necessarily conjoined in speech. In other words, even though many researchers have used the terms ''irony'' and ''sarcasm'' interchangeably [62], these two are not to be mixed with one another as sarcasm has a certain degree of aggression and criticism. This will be addressed in more detail in the next subsection.

In Table 2, we summarize some of the findings related to sarcasm and sarcasm detection from which the research on the automatic detection of sarcasm on social media has benefited.

B. SARCASM AND IRONY

The automatic identification of sarcastic statements has been the subject of several research works conducted by

TABLE 2. Main findings regarding sarcasm, sarcasm detection and sarcasm understanding in previous research works.

Reference	Findings and observations
Jorgensen [59]	• Sarcastic irony is mostly used to complain to or criticize closer ones, who are generally the ones who are hearing the statement. • The degree of social distance or intimacy in the connection between the speaker and the hearer affects the appropriateness of utilizing sarcasm. • The reactions of the hearer (received of the statement) to both sarcastic and non-sarcastic remarks with comparable content were explored, and the difference between the reactions to statements expressing trivial and serious complaints, occurring in conversations between friends of the same sex, were compared. • When complaining, both the hearer and the speaker tend to have negative feelings in both direct and sarcastic complaint conditions. However, sarcasm made it sound more like the speakers is not as serious.
Seckman and Couch [60]	• Sarcasm can fortify and create solidarity within work groups.
Brown and Levinson [63]	• In English, irony may save one's face by allowing one to appear courteous while criticizing satirically. • According to the theory of politeness they developed, being indirect (that is, using sarcasm instead of direct criticism) might be regarded as politeness and therefore not offend the hearer.
Littman and Mey [61]	• Sarcasm and irony are not necessarily conjoined in speech.
Jorgensen [64]	• Because it deals with "echoes" or reminders of previous assertions, verbal irony is very contextual. The previous assertions are either expressly stated in the preceding context or are assumed as shared common knowledge.
Sperber and Wilson [65]	• It is much more common to comment ironically on a failure of some kind than to comment ironically on a success
Giora [11]	• Sarcasm is a mode of indirect negation which requires more than shallow/direct understanding of both the negated and implicated messages.
Gumperz [66]	• Some types of perceptual cues or linguistic signaling can be derived from / recognized in the discussion, leading to interpretations that differ considerably from those that linguistic (syntactic and semantic) clues retrieved from the phrase could lead to. ⇒ Based on this, it is understandable that sarcasm detection from bare texts is considerably more difficult than sarcasm detection in conversations.
Bouazizi and Ohtsuki [47]	• Sarcasm, as employed in social media, is not necessarily a polarity switcher. In several instances, sarcasm conveys a hidden message that is not necessarily the opposite of what has been said.
Kreuz and Glucksberg [67]	• Positive statements can readily be utilized sarcastically since they either contradict common sense or a previous statement (within the context of the conversation). • Only in exceptional instances may negative comments be used sarcastically: using a negative phrase to communicate a good perspective is typically not justified. • When given cues from the speaker, listeners are typically able to detect sarcasm.
Delien et al. [16]	• Detecting sarcasm requires cognitive effort. • Sarcasm identification in spoken interactions is definitely aided by the presence of sarcastic prosody. • The utterance processing required in judging someone else's sarcasm comprehension begins with one's own egocentric perspective. • People (or, at least, the participants) prefer to rely only on cheap, prosodic cues to identify sarcasm whenever these are available.
Rankin et al. [68]	• Sarcasm helps dampthe aggression of the critical comment.
Gibbs [69]	• Sarcasm is detected much faster in instances where explicit echoic mention of some belief, societal norm, or previously stated opinion exists than when the echo is only implicit. • Sarcasm is remembered far better than literal uses of non-sarcastic utterances which express the same meaning.
McDonald [70]	• Sarcasm is more effortful to process than non sarcastic comments. • Inferences regarding the facts of the situation and the speaker's mental state (e.g., attitudes, knowledge, and goals) are critical for understanding sarcasm.

researchers for different purposes. The most common usage of sarcasm detection is to enhance the performance of sentiment analysis systems, which lag behind when sarcasm is employed [9], [71]. In this sense, sarcasm has quite often been confused (and fused) with irony when it comes to their detection in written text. This is because sarcasm is indeed one form of irony. In the Collins online dictionary, irony is defined as *"a subtle form of humor which involves saying things that you do not mean."* In the context introduced above, the definition of irony is no different from that of sarcasm as defined in Section I. However, it is important to emphasize the fact that sarcasm has a criticism aspect and a more "aggressive" attitude added to it. Giora [11] defined sarcasm as a form of "irony that is especially bitter and caustic." Rajadesingan *et al.* [52] suggest that sarcasm is more of "caustic and derisive" type of humor. Quite often than not, the type of irony addressed in the literature in works such as [72]–[74] is the one where the apparent meaning is the opposite of the actual one conveyed by the speaker/writer. The target for identifying this form of irony overlaps with the objective of sarcasm detection as addressed in this paper, as well as others: knowing the original intentions conveyed in the text.

C. TYPES OF SARCASM

While the term "type" might not be the proper way of defining the classification of instances of sarcasm as proposed in the literature, we will be using the term as authors of previous works [9], [47], [52], [75] used it as well.

In [9], [47], sarcasm has been identified as used for 3 main purposes:

- Sarcasm as wit: sarcasm, when used as a wit, has for purpose to be funny. In this context, sarcasm is closer to irony. The person employs some special forms of

speeches, tends to exaggerate, or uses a tone that is different from that when he talks usually to make it easy to recognize.
- Sarcasm as whimper: sarcasm, when used as whimper, has for purpose to show how annoyed or angry the person is, while remaining polite as suggested by the theory of politeness [63].
- Sarcasm as evasion: sarcasm, when used as evasion, has for purpose to avoid giving a clear answer. In other words, rather than criticizing explicitly or saying something that might clearly offend the hearer, the speaker makes use of sarcasm to convey his intentions while remaining polite.

Bharti *et al.* [75] opted for a less complex classification, and defined 7 simple ''types'' of sarcasm. Their definition for sarcasm types is highly correlated with the usage of sarcasm in social media, as some of these types are, by definition, referring to features extracted from social media. The 7 types are:
- T1: The contrast between positive sentiment and negative situation,
- T2: The contrast between negative sentiment and positive situation,
- T3: The use of interjection words at the beginning of posts,
- T4: The contradiction between likes and dislikes,
- T5: A statement contradicting universal facts,
- T6: A statement carrying positive sentiment with its antonym, and
- T7: A statement contradicting time dependent facts.

Similarly, sarcasm has been classified into sub-classes based on how it is employed, rather than what it reflects by Rajadesingan *et al.* [52],
- Sarcasm as a contrast of sentiments: This goes along with many observations made by previous researchers. In this sense, sarcastic utterances use sentimental/emotional words (e.g., *"I love"*) to address or refer to situations that are incompatible with their context (e.g.*"being sick"*).
- Sarcasm as a complex form of expression: This type is based on Rockwell [76]'s observation that there is a small but significant correlation between cognitive complexity and the ability to produce sarcasm.
- Sarcasm as a means of conveying emotion: Here, sarcasm is treated as a mean to convey one's emotions. In other words, in addition to it being a form of aggressive humor [77] or verbal aggression [78], sarcasm is an indirect mean of self-expression as well.
- Sarcasm as a possible function of familiarity: As suggested by [59], [76], sarcasm is more or less used by people towards ones that they are more familiar with. Nevertheless, having a shared knowledge of the language [79] and culture [80] is important to recognize and use sarcasm.
- Sarcasm as a form of written expression: While classically, sarcasm has been addressed as a spoken form of expression, with the exponential growth of social media, people started conveying sarcasm within written texts, by including subtle markers that indicate that the phrase might be sarcastic.

Nevertheless, Camp [12] proposed a different categorization of sarcasm. They suggested that *"different types of sarcasm take different 'scopes', and thereby produce different illocutionary and rhetorical results."* Therefore, they addressed the conventional claim that suggests that sarcasm is straightforward an inversion of the meaning and suggested that sarcastic utterances rather *"pretend to undertake one commitment [...] and they thereby communicate some sort of inversion of this pretended commitment."* They then went ahead and identified 4 types of sarcasm:
- Propositional sarcasm: Here, a proposition is the target of sarcasm and implicit sentiment is conveyed, making the detection of sarcasm in this case quite hard without context. An example for this is *"He's a fantastic guy!"* which, without context might be seen simply as a compliment
- Lexical sarcasm: Here, sarcasm is quite clear and identifiable even without context. An embedded incongruity within the text itself makes the listener/reader identify the sarcasm without resorting to understanding the context itself. For instance in the sentence *"Sam is such a gentleman that no girl wants even to give him a chance!*, the contradiction between the two pieces of information is a clear hint for sarcasm.
- ''Like''-prefixed sarcasm: This also another instance of sarcasm easily detectable and quite often used. ''Like''-prefixed sarcasm is simply sarcasm where the word ''like'' precedes a piece of information that is not correct. An expression such as *"Like I care."* or *"He was like.. I am Bill Gates, aren't I?"* are quite often understood as ''I don't care'' and ''I don't have money,'' respectively.
- Illocutionary sarcasm: This type of sarcasm is quite less commonly used, yet it is the ''ultimate form of sarcasm''. Here, *"the speaker 'makes as if' to undertake a certain speech act S, where S would be appropriate in some counterfactual situation X that contrasts with the current situation Y."* For instance, given a first date between two people (situation Y), where someone acted so poorly, the other person might say *"We should definitely go out again!"* (speech act S) which might be more suited if the first person acted more appropriately (situation X).

Other categorizations of sarcasm and sarcastic statements have been proposed as well. The categorization of sarcasm has been used as the basis for the types of features and subtle markers within texts that could be used to locate sarcastic statements in written text. In the next section, we will discuss in further detail these features and the methods that have been proposed to automatically detect sarcasm.

IV. AUTOMATIC SARCASM DETECTION: METHODS AND APPROACHES

A. RESEARCH WORK ACQUISITION

As stated previously, works on sarcasm detection for sentiment analysis improvement have appeared relatively recently compared to other similar fields. To the best of our knowledge, the first published work to introduce this task was that of Tepperman *et al.* [41] in 2006. Since then, several works have been published, and a large proportion of them was addressing sarcasm detection in social media.

To acquire the different research work done in this field, we queried 3 different search engines for sarcasm detection papers. The 3 search engines we queried are: Google Scholar, IEEE Xplore and ACM Digital Library. We used the following expressions for the search: ''sarcasm detection,'' ''sarcasm recognition,'' and ''sarcasm in social media''.

In total, 264 papers were collected, multiple of which were duplicates or irrelevant to the context of our work have been dismissed. We applied a set of rules to filter out these papers:

1) Only papers whose title or abstract infer directly the idea of sarcasm detection as discussed in this paper are kept.
2) Duplicate papers, or papers from the arXiv (or other preprint websites) whose final versions are found elsewhere are removed.
3) Papers with very poor quality and no significant contribution were removed.
4) We browsed the references of some of the collected papers to find any significant work which we might have missed and included it as well.

The total number of papers directly related to the task of sarcasm detection in texts, in the sense addressed in this paper, is 153.

In the remainder of this we aim to summarize the existing work focusing mainly on the data used, the methods implemented, and the results obtained.

B. SARCASM DATA SETS

Building corpora for automatic sarcasm detection has also been a task investigated deeply as being one of the challenges in the field. This is because sarcasm is very hard to identify and to recognize, even by human annotators, and quite often, the disagreement between annotators is noticeable [17], [81]. In other words, if the annotators have a large disagreement between them in what constitutes sarcasm, it might be hard to build a corpus with well-annotated data, unless the sarcasm within them is clear and indicators in the text are very relevant. That being said, users of social media have invented explicit way to indicate whether what they say is what they mean or not. In particular, in platforms like Twitter, hashtags such as *"#sarcasm"*, *"#irony"* and *"#not"* are still used with sarcastic or ironic statements. By using such key hashtags, to collect tweets, one could collect tweets that were manually ''labeled'' by their own writers as sarcastic. Obviously, similarly to all buzz words and hashtags, these hashtags are quite often abused and/or used by bots to appear in the search results. However, they are still useful to collect an initial set of potentially sarcastic tweets, which needs to be cleaned afterwards.

The hashtag *"#sarcasm"* in particular was used to build several data sets [47], [71] of tweets collected from Twitter. Other works, such as that of Liebrecht *et al.* [45], suggested that *"#not"* is the way to go for sarcastic statement collection However, E. Sulis [82] has shown that the 3 hashtags *"#sarcasm"*, *"#irony"* and *"#not"* are quite different, and should not be confused with each other. Through their experiments on real data, they supported the arguments for the separation between *"#sarcasm"* and *"#irony"*. More interestingly, the hashtag *#not* was qualified as distinct phenomenon, separate from sarcasm and irony in their classical meanings.

1) DATA SETS SOURCE

Throughout the years, several data sets have been built to train and evaluate sarcasm detection approaches. However, despite their diversity, sources of the data are quite few. Following are the top sources of text data sets used for automatic sarcasm detection in the literature:

1) Twitter: Twitter has long been the first option for NLP tasks related to information extraction such as sentiment analysis and automatic sarcasm detection. This is because Twitter is an open platform allowing people to query its API to collect tweets with specific keywords. In particular, as stated above, few hashtags could be very useful to collect sarcastic and sarcasm-related tweets. Several papers have used data sets collected from Twitter such as [83]–[96].
2) Online shops and review websites: Review websites have also been a source of several data sets for tasks such as sentiment analysis. This is because reviews are by definition opinionated texts that show the writer's sentiment toward the product/service he/she is reviewing. Nevertheless, thanks to the scoring system available in many shopping and review websites, no manual annotation is required as the author summarizes his review in a score which can roughly be evaluated as positive if it is high, negative if it is low, and neutral if it is in between. A simple, yet effective way to collect sarcastic texts from such website is to collect texts whose sentiment opposes the score attributed by the author. Works that used data sets collected from Review websites include, among others, [97], [98].
3) Reddit: Reddit is introduced by its creators as a *"is a network of communities based on people's interests."* Reddit has increasingly been a source of information for people with different interests allowing them to rate and discuss any kind of topic, it being a product, service, public figure, etc. While not as straightforward as the previous sources of data, Reddit has the particularity of having more or less a conversation-like structure. This has attracted the interest of researchers

TABLE 3. Sources of data sets used for sarcasm detection.

Source	Works which used data collected from the platform
Twitter	Eisterhold et al. [107], Davidov et al. [17], González-Ibáñez et al. [108], Lunando and Purwarianti [109], Liebrecht et al. [45], Riloff et al. [110], Reyes et al. [72], Ptácek et al. [111], Maynard and Greenwood [71], Barbieri and Saggion [73], Barbieri et al. [46], Liu [112], Tayal et al. [113], Barbieri et al. [114], Bamman and Smith [115], Joshi et al. [19], Rajadesingan et al. [52], Wang et al. [51], Khattri et al. [62], Fersini et al. [116], Bharti et al. [18], Bouazizi and Ohtsuki [9], Muresan [50], Ghosh et al. [117], Bouazizi and Ohtsuki [47], Ghosh and Veale [118], Bouazizi and Ohtsuki [119], Bharti et al. [75], Zhang et al. [120], Sulis [82], Charalampakis et al. [121], Farías et al. [122], Abercrombie and Hovy [48], Poria et al. [53], Joshi et al. [123], Joshi et al. [124], Al-Ghadhban et al. [125], Dharwal et al. [126], Gupta and Yang [127], Mishra et al. [128], Ghosh and Veale [129], Saha et al. [130], Pooja and S. Sarika [131], Mukherjee and Bala [132], Jain et al. [133], Joshi et al. [134], Prasad et al. [135], Ghosh et al. [136], Peled and Reichart [137], Karoui et al. [83], Agrawal and An [84], Ren et al. [85], Parmar et al. [86], Kannangara [87], Samonte et al. [88], Bohra et al. [89], Liu et al. [90], Cai et al. [91], Son et al. [92], Tao et al. [93], Bharti et al. [94], Jain et al. [95], Ren et al. [96], H. Elgabry et al. [138], D. Faraj et al. [139], H. Nayel et al. [140], M. Shrivastava and S. Kumar [141], A. Mahdaouy et al. [142], C. Lou et al. [143], F. Husain et al. [144], M. S. Razali et al. [145], F. Yao et al. [146], X. Guo et al. [147], A. Kamal et al. [148], Y. Du et al. [149].
Reddit	Joshi et al. [123], Ghosh et al. [136], Khodak et al. [99], Agrawal and An [84], Ghaeini et al. [150], Liu et al. [90], Tao et al. [93], Ren et al. [96], Justo et al. [100], Hazarika et al. [101], Y. Zhang et al. [151], C. Lou et al. [143], Y. Du et al. [149].
Amazon	Davidov et al. [17], Filatova [81], Buschmeier et al. [97], Liu [112], Filatova [98].
Others	Eisterhold et al. [107], Tepperman et al. [41], Kreuz and Caucci [42], Lukin and Walker [44], Rakov and Rosenberg [152], Reyes and Rosso [74], Dave and Desai [102], Joshi et al. [49], Joshi et al. [103], Mishra et al. [128], Suhaimin et al. [104], Joshi et al. [134], Agrawal and An [84], Das and Clark [105], Das [106], Liu et al. [90], Y. Zhang et al. [151], M. Bedi et al. [153], C. Lou et al. [143], A. Kumar et al. [154], Y. Wu et al. [155], Z. Wen et al. [156].

as it offers more than just one-sided opinions of people, but rather more detailed discussions where people can ask for clarifications or argue against an opinion, etc. Fewer works have used Reddit as a source of their data. This is because, unlike the previous sources, data collected from this platform require manual annotation. It is worth mentioning, however, is that most of the works used the data set offered by Khodak *et al.* [99], being the first of its kind, and containing 1.3 million sarcastic statements. Works that used data sets collected from Reddit include, among others, [96], [100], [101].

4) Others: In addition to the previously mentioned sources of data, few works have used data such as Facebook, TV series transcripts (e.g. ''Friends'', ''Daria''), google books-extracted texts, online forums, blogs, etc. [102]–[106]

In Table 3, we describe which works used these sources of data in their work. As can be observed, most of the works used Twitter as their primary source of data used for sarcasm detection. In addition, in Table 4, we give examples of some of the data sets available online which have been used in these works.

2) LANGUAGE

In terms of language used in the data collected, most of the work presented above dealt with English texts. Not only is English the language used the most on Internet, but also the tools for text processing and feature extraction are more mature for English than for other languages. Part-of-Speech tagging, lemmatization, stemming, automatic summarization, named entity recognition and relationship extraction are some of the basic NLP tasks that have reached impressive performance for English, while still struggling in other languages, in particular non-Latin derived ones.

In the context of sarcasm detection, works that addressed languages other than English are quite few. Latin descendant languages that have been addressed include, but are not limited to French [83], Italian [46], [83], Dutch [45], Czech [111], etc.

Non-Latin descendant languages also have been addressed in few works. These include the following ones. Lunando and Purwarianti [109] employed translated SentiStrength [161] to extract sentiment-related features from Indonesian text to perform sarcasm detection. Liu [112] introduced a set of features specifically for detecting sarcasm in social media for Chinese. Charalampakis *et al.* [121] compared supervised techniques with unsupervised ones for sarcasm detection in Greek. Dave and Desai [102] studied different classification techniques for sarcasm detection and experimented on Hindi blog reviews. Similarly, Bharti *et al.* [94] and Jain *et al.* [95] targeted Hindi in their work on sarcasm detection on Twitter. Al-Ghadhban *et al.* addressed the problem of sarcasm detection for Arabic, and evaluated their approach on a data set collected from Twitter and manually labeled. Suhaimin *et al.* [104] performed the same task for Malay on posts collected from Facebook. Samonte *et al.* [88] performed sentence-level sarcasm detection on tweets collected about government, politics, weather, social media, and public transportation for English and Filipino, and showed a significant difference in the results of their experiments.

C. SARCASM DETECTION: METHODS AND APPROACHES

The detection of sarcasm in the literature has mostly taken the form of a classification problem, with very few exceptions. The idea is roughly running a classification task on a set of

TABLE 4. Example of data sets available online.

Data set name	Domain	Reference	Description
SASI-AM	Amazon reviews	[43]	. 67 sarcastic texts . 113 non-sarcastic texts
SASI-TW	Twitter	[43]	. 73 sarcastic texts . 107 non-sarcastic texts
RILOFF	Twitter	[110]	. 112 sarcastic texts . 498 non-sarcastic texts
ELECT	Twitter	[157]	. 938 sarcastic texts . 938 non-sarcastic texts
FILATOVA	Reviews	[81]	. 437 sarcastic texts . 437 non-sarcastic texts
IAC-SARC2	Reddit posts	[158]	. 4692 sarcastic texts . 4692 non-sarcastic texts
SARC	Reddit posts	[99]	. 1.34M sarcastic texts . 532M non-sarcastic texts
IAC-Subset [Walker et al.]	Reddit posts	[159]	. 753 sarcastic texts . 13371 non-sarcastic texts
BOUAZIZI	Twitter	[47], [119]	. 3500 sarcastic texts . 3500 non-sarcastic texts
PTACEK 2014	Twitter	[111]	. 50,000 sarcastic texts . 50,000 non-sarcastic texts
ArSARCASM-V2	Twitter	[160]	. 10,380 sarcastic texts . 2,168 non-sarcastic texts

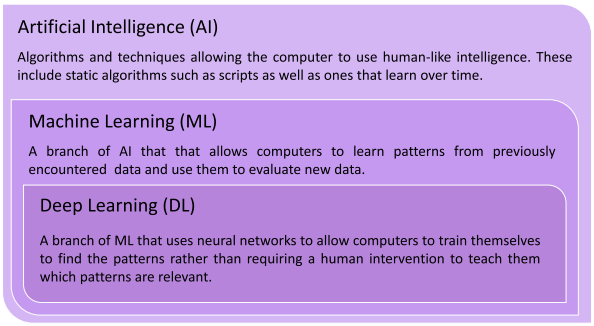


FIGURE 2. Relation between artificial intelligence, machine learning and deep learning.

texts and identify which ones are sarcastic and which ones are not. That being the case, Artificial Intelligence (AI) has been the way to go to perform such a task.

Roughly speaking, as shown in Figure 2, AI could be thought of as the use of computers to mimic the human brain behavior in performing certain tasks. Machine Learning (ML) is a particular type of AI which, given a set of manually labeled data, and a set of rules to extract patterns from these data, could learn how to deal with new unseen data reliably. Deep Learning (DL) is a branch of ML in which the learning of patterns and identifying which are relevant is automatized and left for the computer itself to do.

In Figure 3 we show the overall flowchart of use of ML and DL for classification. Basically, a manually annotated data set (a set of objects –i.e., texts– alongside with the class they belong to) is given to the ML or DL algorithm. This data set is usually referred to as the training set. The algorithm extracts specific patterns from these objects that allow it to recognize their classes. The process of learning these patterns and the relations between them is referred to as the training phase. Upon training, the model is given an unknown object (i.e., does not belong to the training set), and is asked to identify its class by extracting the same features and comparing them to its knowledge. A good model should predict unseen objects with high accuracy. The main difference between ML and DL is the pattern extraction procedure itself. In ML, a human should teach the machine which features to extract from the input training data, upon which the machine builds its internal rules to recognize objects from these features. In DL, the human intervention is limited to the "design" of the neural network and its hyper-parameters. The network learns which features are relevant and which are not all by itself.

In the context of this paper, ML, and recently DL, have been dominantly the ultimate method to detect sarcasm. Nevertheless, other approaches that do not use supervised learning have been proposed. In the rest of this subsection, we summarize the methods used and approaches proposed.

1) RULE-BASED APPROACHES

Rule-based approaches are approaches that define a set of rules according to which a statement is judged as sarcastic or not. For instance, Maynard and Greenwood [71] have used hashtags to identify sarcastic statements. In part of their works relied on explicit hashtags such as *"#sarcasm"* and *"#Irony"* to identify sarcasm. Nevertheless, they also investigated in more details more complex hashtags: they proposed an approach to re-tokenize the hashtags and use the information extracted from them to identify if a statement is sarcastic or not. For example, in the text *"You are more than welcome! #notreally"*, the hashtag is transformed into the expression it says *"not really"*, which contradicts the content of the tweet. Therefore, it could be concluded that this tweet is indeed sarcastic. Riloff *et al.* [110] proposed a method to

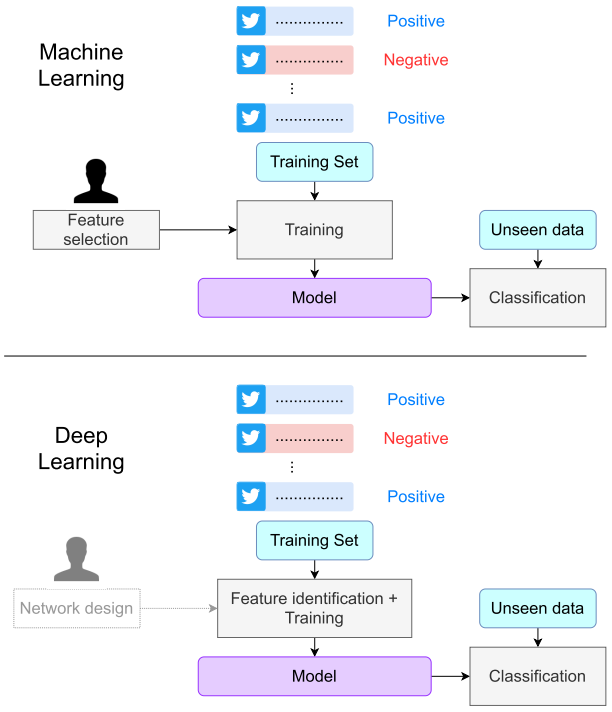


FIGURE 3. Machine learning and deep learning training and classification procedures.

detect a particular type of sarcasm in which the author uses a positive sentiment to describe his feelings towards a negative situation. Their method relies on a bootstrapping algorithm that starts with the seed word "love" and a set of sarcastic tweets to build a set of positive sentimental words and a of set negative situations, which they used to judge when there is sarcasm and when there is not. Other works such as [47] iterated further on the idea of contradiction between positive and negative components within a piece of text to decide whether or not it is sarcastic.

2) MACHINE LEARNING APPROACHES

Conventional machine learning, in particular, has been intensively explored. Most of the existing works up-to-date followed the same pattern: extract a set of features and use machine learning algorithms such as Naive Bayes [162], Support Vector Machine (SVM) [163], Maximum Entroy [164]. Features are manually engineered and carefully chosen to highlight any sarcastic-related information.

In Table 5, we summarize the most common types of features that have been used to train such classifiers. These types of features are explained in more details below.

a: LEXICAL FEATURES

Lexical features are simply features that use the basic components of a given text, such as n-grams, hashtags, etc. These are the most basic types of features, yet they are employed more than any other type of features. They have been used not only in sarcasm detection, but also sentiment analysis, hate speech detection. Lexical features have been used in most of the existing work [42], [47], [108], and have given promising results.

b: PRAGMATIC FEATURES

Pragmatic features are features that exploit features other than the text itself, such as emoticons, user mentions and some hashtags. Pragmatic features are mostly used in social media-collected data sets, and have proven to be very efficient in detecting sarcasm on such data sets. Pragmatic features have been used in several works such as [47], [72], [108], [119], [121]

c: PATTERN FEATURES

Pattern features are features that exploit the repetitiveness in user-text when expressing sarcasm. Common expressions showing sarcasm have been widely used (e.g., *"I love it when + negative clause"*). Multiple works [17], [44], [47] have used this family of features, and patterns are built either by exploiting the frequency of usage of words, their grammatical functions or pre-built expressions.

d: CONTEXTUAL FEATURES

Contextual features are features that exploit the context of the text, in addition to its content. Contextual features require a knowledge beyond the text itself. For instance, if a given message (typically a tweet in the case of Twitter data sets) is a reply to another one, the knowledge of the content of the original message could help identify sarcasm more accurately. Nevertheless, the knowledge of at least a sub-part of the content of the data set itself is required. Several approaches [41], [52], [74], [115] rely on the understanding of what makes a situation negative or positive, which requires either manual effort by the annotators or building a system to collect a data set of such negative situations.

e: SENTIMENT AND EMOTION FEATURES

Sentimental features are simply the same kind of features used typically in sentiment analysis classification. They include features related to the usage of sentimental words (i.e., positive and negative words), exaggeration of expressing emotions and contradiction between emotional words within the same sentence. Works that used this type of features include [73], [74].

f: BEHAVIORAL FEATURES

Behavioral features [52], [131] are features that make use of the understanding of the behavior of the Internet user to identify his typical behavior in normal situation and when he is employing sarcasm. Such understanding is built over a certain period of time, and is used to identify when sarcasm is employed and when it is not. Behavioral features could be seen as the observation over time of other types of features. This is because these other features change over time is the fundamental information used.

TABLE 5. Common types of features used in machine learning approaches.

Features	Definition
Lexical features	Features that use the basic components of a given text, such as n-grams, interjections, hashtags, etc.
Pragmatic features	Pragmatic features are features that exploit features other than the text itself, such as emoticons, user mentions and some hashtags.
Pattern-based features	Pattern features are features that exploit the repetitiveness in user-text when expressing sarcasm.
Contextual features	Contextual features are features that exploit the context of the text, in addition to its content.
Sentiment and emotion features	Sentimental features include features related to the usage of sentimental words (i.e., positive and negative words), exageration of experssing emotions, etc.
Behavioral features	Behavioral features are features that make use of the understanding of the behavior of the Internet user to identify his typical behavior in normal situation and when he is employing sarcasm.
Syntactic features	Syntactic features are feature related to the arrangement of words and phrases to create well-formed sentences.
Metaphoric features	A metaphor is a figure of speech in which a word/expression is used to describe an object, event or idea where it is not literally applicable.
Hyperbole features	Hypergole features are similar in nature to metaphoric features. They use extreme comparisons and exagerations to make a point or show emphasis.
Sociolinguistic features	Sociolinguistic are user-related features, which focus on his information rather than the text's extracted ones.
Prosodic and acoustic features	Proposodic features are features focusing on the elements of speech in a sentence as a whole such as the intonation, tone, stress and rhythm. Such features are usually used in vocal speech.
Punctuation features	Punctuation features are simply features related to the use of marks in the text. They can reveal some sort of intention the user might intend to convey.
Semantic features	Semantic features are ones particular to languages, as they relate to the meaning in the language or the logic behind common expressions or phrases, etc.
Rhetorical feature	Rhetorical feature are specific to some languages such as East-Asian ones. They includes extreme nouns, adjectives or adverbs, as well as titles of degrees and honorifics.
Image/video features	These include indicators extracted from content other than the text itself such as images and videos posted with the text.
Personality features	Personality features are features related to the behavior and thought patterns of people.
Stylistic features	Stylistic features refer to features related to the idiolect and authorship styles of users, which is reflected in their writing
Idiosyncratic features	Idiosyncratic features refer to very strange and unsual expressions, metaphors or comparisons that are not usually employed in conversations.
Embeddings	Embeddings are numeric representation of words which are typically generated by training deep neural networks on large corpora to identify relation between the words of a given language.

g: SYNTACTIC FEATURES

Syntactic features [104], [111], [112] are feature related to the arrangement of words and phrases to create well-formed sentences. Several of the so-called memes are used to express sarcasm, and use intentionally grammatically wrong sentences (e.g., *"All your base are now belong to us"*) to mock others' lack of knowledge, their bad language, and in general emphasize any bad quality. Syntactic features are highly correlated with the idea of Part of Speech (PoS) tags, as typical sentences follow certain patterns of PoS tags.

h: METAPHORIC FEATURES

A metaphor is a figure of speech in which a word/expression is used to describe an object, event or idea where it is not literally applicable. For example, using the expression ''lone wolf'' to describe introverts is a common metaphor. Metaphoric features [47] are ones that make use of metaphor to express sarcasm. This also includes the use of commonly agreed on knowledge to ridicule something.

i: HYPERBOLE FEATURES

Hyperbole features [49], [97] are similar in nature to metaphoric features. They use extreme comparisons and exaggerations to make a point or show emphasis. In the context of sarcasm, such features are employed to tell the opposite of something quite obvious. For example, one might refer to an obese person by saying *"He's as skinny as a toothpick."*

j: SOCIOLINGUISTIC FEATURES

Sociolinguistic features are user-related features, which focus on his information rather than the text's extracted ones. They include for example the age, the gender, etc. Some works that used sociolinguistic features include [115] and [48].

k: PROSODIC FEATURES

Proposodic features [41], [104], [152] are features focusing on the elements of speech in a sentence as a whole such as the intonation, tone, stress and rhythm. Such features are usually used in vocal speech. However, they usually translate into other forms in written text, such as the repetition of a cerain vowl or the use of capitalization to convey some intonation, etc.

l: PUNCTUATION FEATURES

Similar to how prosodic features translate in some particular use of capitalizations, etc., punctuation can also reveal some sort of intention the user might intend to convey. For instance, the excessive use of exclamations marks (e.g., *"That is amazing!!!!!"*), or question marks (e.g., *"Oh really I dindn't know!!"*) could reveal the sarcastc aspect of the user. Several works have used punctuation features, along with others, for

sarcasm detection. They include [100], [113], [116], [122], etc.

m: SEMANTIC FEATURES

Semantic features [46], [112] are ones particular to languages, as they relate to the meaning in the language or the logic behind common expressions or phrases, etc.

n: RHETORICAL FEATURES

Rhetorical features [112] are specific to some languages such as East-Asian ones (e.g., Chinese and Japaneses). They include extreme nouns, adjectives or adverbs, as well as titles of degrees and honorifics.

o: PERSONALITY FEATURES

Personality features [53], [82] are features related to the behavior and thought patterns of people. Pre-trained models on the automatic detection of personality traits (Big 5 for example) could be applied to one's posts/comments/tweets to extract his personality traits, which in return could be used as features to identify sarcasm.

p: STYLISTIC FEATURES

People possess their own idiolect and authorship styles, which is reflected in their writing. These styles are generally affected by attributes such as gender, diction, syntactic influences, etc. Works that used this type of features include [114], etc.

q: IDIOSYNCRATIC FEATURES

The term ''Idiosyncrasy'' refers to an odd habit or a peculiar way of behavior/thought. It is commonly used to express eccentricity or more generally strange and weird attributes. In the context of linguistics, the term could refer to very strange and unusual expressions, metaphors or comparisons that are not usually employed in conversations. If employed, they intend to bring a particular meaning or cue to the conversation, including sarcastic cues [104], [165].

r: EMBEDDINGS

Word embeddings [90], [92], [166] are numeric representations of words and expression which were attributed through advanced techniques such as ''skip-gram'' [167]. While the meaning of these numeral values are hidden and not directly interpretable by humans, neural networks, in particular, make use of such representation to perform complex tasks related to NLP.

Classifier-wise, most of the works have used the following classifiers:

- Support Vector Machine: SVM is a robust prediction method based on statistical learning frameworks or VC (Vapnik–Chervonenkis) theory [168]. In an SVM,training examples are mapped into points in a multidimentional space with the objective of maximizing the gap between the two classes (for the case of binary classification). In the context of sarcasm detection, SVM is amongst the most used algorithms and performing the best in terms of classification accuracy and precision. LibSVM [169] is probably the most used implementation of SMV.
- Naive Bayes: Naive Bayes is probably the simplest, yet one of the top performing classifiers in NLP-related classification tasks. A Naive Bayes classifier is a probabilistic classifier based on the idea of applying Bayes' theorem with strong independence assumptions between the features.
- Maximum entropy: A Maximum Entropy classifier [170] is a discriminative classifier based on the statement that suggests that the probability distribution which best represents the current state of knowledge is the one with the largest entropy. This classifier is widely used in NLP problems, including sarcasm detection.
- Logistic Regression: Logistic regression [171] is basically a statistical model that models a binary dependent variable, and thus in its core is not a classification operation. However, transforming it into a binary classifier could be done by defining some threshold for the continuous output, below which the input is judged as belonging to one class, and above which the input is judged as belonging to another class. Regression in the context of NLP has no clear meaning. However, Logistic Regression-based classifiers have shown great potentials.
- Random Forest: Random Forest classifiers [172] are a common type of decision tree-based classifier used in a variety of tasks. In Random Forests, multiple decision trees are constructed, and an ensemble method is applied to them. Decision trees have shown great potentials in tasks related to NLP, and are among the top performing classifiers.
- k-nearest neighbors (kNN): The k-nearest neighbors classifier is a non-parametric classification method in which the input consists of the k closest training examples in data set and the output is a class membership. A definition of distance is required to identify what constitute ''near'' neighbor. Depending on the value of k and the weighting used to favor closer neighbors, the classification is basically done by averaging the weights of the classes of the k closest examples from the training set and picking the maximum one.

Other classifiers used include decision trees, SMV-Hidden Markov Models (SVM-HMM), Gradient boosting [173], or Searn [174], etc.

In Tables 6 and 7, we show a summary of the works which used the sets of features introduced above. In Table 8, we show a summary of the works which used the machine learning algorithms described above.

3) DEEP LEARNING APPROACHES

With the recent advances on the filed of Deep Learning (DL), mainly the contributions of Lecun *et al.* [176] Hinton *et al.* [177] and Krizhevsky *et al.* [178], it became

TABLE 6. Features used in the machine learning-based approaches (Part 1).

Year	Authors	Lexical features	Pragmatic features	Pattern-based features	Contextual features	Sentiment features	Behavioral features	Syntactic features	Metaphoric features	Hyperbole features	Sociolinguistic features	Prosodic and acoustic features	Punctuation features	Semantic features	Rhetorical feature	Personality features	Stylistic features	Idiosyncratic features	Embeddings	Others
2006	Tepperman et al. [41]				X							X								
2007	Kreuz and Caucci [42]	X																		
2010	Davidov et al. [17]			X																
2011	González-Ibáñez et al. [108]	X	X																	
2013	Lunando and Purwarianti [109]	X				X														
2013	Lukin and Walker [44]			X																
2013	Liebrecht et al. [45]																			
2013	Rakov and Rosenberg [152]	X										X								
2013	Reyes et al. [72]	X	X			X		X						X						
2014	Justo et al. [100]	X											X							
2014	Ptácek et al. [111]	X		X				X					X							X
2014	Barbieri and Saggion [73]	X				X							X							
2014	Reyes and Rosso [74]	X	X		X	X														
2014	Buschmeier et al. [97]	X	X			X				X			X							
2014	Barbieri et al. [46]	X				X								X						
2014	Liu [112]	X						X					X	X	X					
2014	Tayal et al. [113]	X	X			X							X							
2015	Barbieri et al. [114]	X				X								X			X			
2015	Bamman and Smith [115]				X	X		X			X		X							
2015	Joshi et al. [19]	X	X			X														
2015	Rajadesingan et al. [52]	X			X	X	X							X						
2015	Wang et al. [51]	X			X															
2015	Khattri et al. [62]	X			X	X														
2015	Fersini et al. [116]	X	X																	
2015	Bharti et al. [18]	X	X							X			X							
2015	Bouazizi and Ohtsuki [9]	X	X	X		X		X		X			X	X						
2015	Muresan [50]	X	X			X														
2015	Ghosh et al. [117]				X														X	
2016	Dave and Desai [102]	X	X		X															
2016	Bouazizi and Ohtsuki [47]	X	X	X		X		X	X	X			X	X						
2016	Joshi et al. [49]	X	X		X	X				X			X							
2016	Bouazizi and Ohtsuki [119]	X	X	X		X		X		X			X	X						
2016	Bharti et al. [75]	X	X							X			X							
2016	Zhang et al. [120]	X																	X	
2016	Sulis [82]	X				X										X				
2016	Charalampakis et al. [121]	X	X			X														
2016	Farías et al. [122]	X	X			X							X	X						
2016	Abercrombie and Hovy [48]	X									X		X							
2016	Poria et al. [53]					X										X			X	
2016	Joshi et al. [123]	X																		
2016	Joshi et al. [103]																		X	
2017	Al-Ghadhban et al. [125]		X			X							X							
2017	Dharwal et al. [126]	X			X	X														
2017	Gupta and Yang [127]	X				X													X	
2017	Mishra et al. [128]	X			X															
2017	Ghosh and Veale [129]				X														X	
2017	Suhaimin et al. [104]	X	X					X				X						X		
2017	Saha et al. [130]	X																		
2017	Pooja and Sarika [131]			X		X	X	X					X							
2017	Mukherjee and Bala [132]	X		X										X						
2017	Jain et al. [133]	X	X			X														
2017	Filatova [98]	X				X														
2017	Joshi et al. [134]				X	X								X					X	
2017	Prasad et al. [135]	X				X							X	X						
2017	Ghosh et al. [136]	X		X		X							X							
2017	Karoui et al. [83]				X	X									X					
2017	Bharti et al. [94]				X	X														

TABLE 7. Features used in the machine learning-based approaches (Part 2).

Year	Authors	Lexical features	Pragmatic features	Pattern-based features	Contextual features	Sentiment features	Behavioral features	Syntactic features	Metaphoric features	Hyperbole features	Sociolinguistic features	Prosodic and acoustic features	Punctuation features	Semantic features	Rhetorical feature	Personality features	Stylistic features	Idiosyncratic features	Embeddings	Others
2018	Khodak et al. [99]	X																		
2018	Agrawal and A. An [84]					X													X	
2018	Ghaeini et al. [150]																		X	
2018	Hazarika et al. [101]															X	X		X	
2018	Ren et al. [85]				X														X	
2018	Parmar et al. [86]	X				X				X										
2018	Kannangara [87]	X			X			X												
2018	Das and Clark [105]		X			X														X
2018	Samonte et al. [88]	X	X							X										
2018	Bohra et al. [89]	X	X			X							X							
2018	Das [106]		X			X														X
2019	Liu et al. [90]							X					X							X
2019	Di Gangi et al. [166]													X						X
2019	Son et al. [92]												X							X
2019	Tao et al. [93]	X																	X	
2019	Castro et al. [175]											X							X	
2020	Jain et al. [95]		X		X	X													X	
2020	Ren et al. [96]					X								X						
2021	H. Elgabry et al. [138]																		X	
2021	Y. Zhang et al. [151]																		X	X
2021	D. Faraj et al. [139]																		X	
2021	H. Nayel et al. [140]	X																		
2021	M. Shrivastava et al. [141]		X																X	
2021	M. Bedi et al. [153]	X										X						X	X	
2021	A. Mahdaouy et al. [142]																		X	
2021	C. Lou et al. [143]	X			X														X	X
2021	A. Kumar et al. [154]	X			X															
2021	Y. Wu et al. [155]	X			X							X							X	X
2021	F. Husain et al. [144]																		X	
2021	M. S. Razali et al. [145]	X			X	X				X										
2021	F. Yao et al. [146]	X			X						X								X	X
2021	X. Guo et al. [147]																		X	X
2022	A. Kamal et al. [148]																		X	X
2022	Y. Du et al. [149]				X	X								X						X
2022	Z. Wen et al. [156]													X					X	X

possible to train big Neural Networks (NN) in a reasonable amount of time while keeping the training converge in most of the time. This has led to more interest towards the use of NN in almost all learning-related fields, and DL has replaced older techniques, including conventional ML, in tasks varying from image recognition [179] to natural language translation [180], or even text generation [181] and style transfer [182].

Nevertheless, text mining has been one of the main domains that have profited from this technique. In particular, for the task of automatic sarcasm detection, several works have been introduced in the past few years that used deep learning to perform sarcasm detection. While the common stream suggests using Recurrent Neural Networks (RNN) -based techniques for text processing and classification as text is usually considered a sequence of words, few works have used more conventional approaches that use CNN on its essence to perform the classification. In Table 9, we present a brief summary of these works.

The main families of DL approaches are the following:

- Convolutional Neural Networks (CNNs): Convolutions are the basics of the modern neural network architectures. CNNs are deep neural networks based on convolutions, most commonly applied in computer vision. In the context of text classification, texts are converted into numeric matrices, to which 1-D convolution is applied to perform the classification. In addition to the final dense layer which generates the class probabilities, another dense layer connects it to the flattened output of the convolutional layers.
- Long Short Term Memory (LSTM) Networks: LSTMs are a specific type of Recurrent Neural Networks.

TABLE 8. Common machine learning algorithms used for sarcasm detection.

Classifier	Works which used the classifier
Random Forest	Bouazizi and Ohtsuki [9], [47], [119], Sulis [82], Charalampakis et al. [121], Pooja and Sarika [131], Jain et al. [133], Filatova [98], Prasad et al. [135], Das and Clark [105], Das [106], Di Gangi et al. [166], H. Elgabry et al. [138], A. Kumar et al. [154].
Supprov Vector machine	Ptácek et al. [111], Buschmeier et al. [97], Liu [112], Joshi et al. [19], Rajadesingan et al. [52], Wang et al. [51], Fersini et al. [116], Bouazizi and Ohtsuki [9], Muresan [50], Ghosh et al. [117], Dave and Desai [102], Bouazizi and Ohtsuki [47], Joshi et al. [49], Bouazizi and Ohtsuki [119], Sulis [82], Charalampakis et al. [121], Farías et al. [122], Joshi et al. [123], Joshi et al. [103], Dharwal et al. [126], Gupta and Yang [127], Mishra et al. [128], Suhaimin et al. [104], Saha et al. [130], Pooja and Sarika [131], Filatova [98], Ghosh et al. [136], Ren et al. [85], Das and Clark [105], Samonte et al. [88], Das [106], Di Gangi et al. [166], Castro et al. [175], H. Nayel et al. [140], A. Kumar et al. [154], M. S. Razali et al. [145].
k-NN	Davidov et al. [17], Pooja and Sarika [131], Filatova [98], M. S. Razali et al. [145].
Logistic regression	González-Ibáñez et al. [108], Bamman and Smith [115], Rajadesingan et al. [52], Muresan [50], Sulis [82], Abercrombie and Hovy [48], Dharwal et al. [126], Mishra et al. [128], Jain et al. [133], Filatova [98], Prasad et al. [135], Di Gangi et al. [166], H. Nayel et al. [140], M. S. Razali et al. [145].
Naïve Bayes	Reyes et al. [72], Justo et al. [100], Buschmeier et al. [97], Liu [112], Bouazizi and Ohtsuki [9], Muresan [50], Bouazizi and Ohtsuki [47], Bouazizi and Ohtsuki [119], Sulis [82], Charalampakis et al. [121], Farías et al. [122], Al-Ghadhban et al. [125], Mishra et al. [128], Saha et al. [130], Mukherjee and Bala [132], Jain et al. [133], Prasad et al. [135], Das and Clark [105], Samonte et al. [88], Das [106], H. Nayel et al. [140].
Decision Tree	Reyes et al. [72], Barbieri and Saggion [73], Barbieri et al. [46], Barbieri et al. [114], Rajadesingan et al. [52], Fersini et al. [116], Sulis [82], Charalampakis et al. [121], Farías et al. [122], Filatova [98], Prasad et al. [135], M. S. Razali et al. [145].
Maximum Entropy	Ptácek et al. [111], Liu [112], Pooja and Sarika [131], Samonte et al. [88]
Others	González-Ibáñez et al. [108], Liebrecht et al. [45], Rakov and Rosenberg [152], Wang et al. [51], Fersini et al. [116], Joshi et al. [49], Charalampakis et al. [121], Mukherjee and Bala [132], Jain et al. [133], Filatova [98], Prasad et al. [135], Prasad et al. [135], Das and Clark [105], Das [106], Di Gangi et al. [166], H. Nayel et al. [140], A. Kumar et al. [154].

TABLE 9. Deep learning-based approaches for sarcasm detection.

Year	Authors	Approach
2016	Ghosh and Veale [118]	Conv-LSTM
2016	Zhang et al. [120]	Bi-GRNN
2016	Poria et al. [53]	CNN
2017	Ghosh and Veale [129]	ConvLSTM
2017	Ghosh et al. [136]	LSTM + Attention
2018	Agrawal and An [84]	LSTM
2018	Ghaeini et al. [150]	Bi-LSTM + Attention
2018	Hazarika et al. [101]	CNN
2019	Cai et al. [91]	Bi-LSTM + Attention
2019	Son et al. [92]	Bi-LSTM + Attention
2019	Tao et al. [93]	Bi-LSTM
2020	Jain et al. [95]	ConvLSTM
2020	Ren et al. [96]	ConvLSTM + Attention
2021	D. Faraj et al. [139]	Transformers
2021	M. Shrivastava et al. [141]	Transformers
2021	M. Bedi et al. [153]	LSTM + Attention
2021	A. Mahdaouy et al. [142]	Attention
2021	C. Lou et al. [143]	LSTM + Attention
2021	Y. Wu et al. [155]	CNN + Transformers
2021	F. Husain et al. [144]	Transformers
2021	M. S. Razali et al. [145]	CNN
2021	F. Yao et al. [146]	CNN + Bi-LSTM
2021	X. Guo et al. [147]	Transformers
2022	A. Kamal et al. [148]	Bi-RNN + Attention
2022	Y. Du et al. [149]	CNN/Bi-LSTM + Attention
2022	Z. Wen et al. [156]	GRU + Attention

LSTMs were invented to solve the vanishing gradient problem that were often occurring when training RNNs, in which long-term previous components have an exponentially decreasing effect on later components. This is because they can learn order dependence in sequence prediction problems, including long-term dependence.

- Bidirectional-LSTM (Bi-LSTM): Bi-LSTMs are a particular type of LSTM in which two "independent" LSTMs are put together each processing the time-dependent items (i.e., words in our case) in both chronological order, and backwards one. This allows the networks to have both backward and forward information about the sequence at every time step.
- Attention Networks: attention networks are basically an iteration over classic RNNs and LSTMs, in which the encoder-decoder architecture is "freed" from the fixed length internal representation. While a classic LSTM forces the encoder to take into account all the previous items, an attention model allows it to focus only on certain inputs in the input sequence for each output item.

With regards to this current work, we explore in the next sections two main streams of approaches: ones that employ conventional machine learning trained with several feature sets including patterns and ones that use deep learning (LSTM and LSTM with attention). We also study sarcastic statements on 3 different platforms: Twitter, Reddit, and some News websites.

D. REPORTED RESULTS

In the literature, several Key Performance Indicators (KPIs) have been used to evaluate the efficiency of the proposed approaches. Following is a list of the most commonly used ones and how they are measured.

- **True Positive Rate (TPR):** refers to the ratio of correctly classified elements over the entire input.
- **Precision:** refers to the ratio of relevant elements over the retrieved instances.

- **Recall:** refers to the ratio of relevant elements that are retrieved over the total amount of relevant elements.
- **F_1-score:** which is a measure that combines both precision and recall, used usually to compare different approaches, and defined as follows:

$$F_1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (1)$$

The F_1-score is an instance of a larger family of scores referred to as F_β where β is a coefficient given the precision to change its weight compared to that of the recall. In general, F_β is defined as follows:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

The F_1-score is sometimes referred to as the F1-score, F-measure or simply, yet somewhat imprecisely, F1, all of which are agreed on.

- **Accuracy**: across an entire data set, accuracy is similar to the TPR, as it measures the ratio of correctly classified instances over the entire set of input instances.
- **Area under the curve (AUC):** the AUC measures the ratio of the area under the Receiver Operating Characteristic (ROC) curve to the total TPR to False Positive Rate (FPR) area. The ROC curve itself is a graph that shows the performance of a classification model at different classification thresholds.

In Table 10, we show some of the results reported in some of the works existing in the literature. It is a bit inaccurate though to directly judge these works by comparing them to one another, since the data sets that were used are quite different, and the results reported in these are highly correlated to their respective data sets. We limit the reported results to ones run on Twitter data sets. Note that for works that reported results for multiple approaches or on multiple data sets, we limit the shown results to the best reported ones.

E. SUMMARY

In Figure 4, we show a summary of the most relevant work proposed over the last decade or so regarding the automation of sarcasm detection. As can be seen, most of the work was performed on data sets collected from Twitter. In addition, unsurprisingly, English is the language on which most of the studies have been performed. Other language have been addressed in very few works in the literature. Nevertheless, while the vast majority of these works have used machine learning (SVM, Random Forest, and Naive Bayes), recent years have seen an increase in the use of DL-based methods, in particular, CNN, LSTM and transformers. We can clearly observe that DL approaches started to appear mostly after 2015, with the advances that this field has been subject to, in particular ones related to NLP.

V. CHALLENGES AND OPEN RESEARCH PROBLEMS

Below are some of the challenges related to the task of sarcasm detection that are yet to be investigated further and present open problems for research.

A. ANNOTATION OF THE DATA

As previously stated, creating an annotated sarcasm data set remains a major challenge. Despite the efforts made by several researchers such as Khodak *et al.* [99], Riloff *et al.* [110], Filatova [81] and Bouazizi and Ohtsuki [47], a well-elaborate data set is yet to be built: On the one hand, sarcasm is in many cases contextual. This context comes at different levels: previous messages in the conversation that led to the sarcastic statement, the relation between the speaker and the listener, the fact (if it exits) that the sarcasm negates, etc. On the other hand, sarcasm, as suggested by Davidov *et al.* [17] and Bouazizi and Ohtsuki [47] comes with different ''intensities'' or different levels. Under such assumption or hypothesis, when annotating a text, it might be important to attribute a score or an intensity level to each piece of text.

On a related topic, comparing works to one another is quite hard when each is experimenting with a self-made data set. Making a standard data set for sarcasm detection evaluation is very important. Here, a competition such as PAN at CLEF,[3] has proven to be a good reference allowing researchers to compete and create a robust benchmark for future works in multiple NLP tasks such as hate speech detection, author profiling, etc. Similar competitions for sarcasm detection can be a start point for such a robust reference for benchmarking and evaluating future works.

B. SARCASM AND IRONY

In several works in the literature, sarcasm was confused, intentionally or unthinkingly, with irony. However, as stated above, Littman and Mey [61] suggested that sarcasm and irony are not necessarily conjoined in speech. Distinguishing one from the other is a quite hard task, and is by far more challenging than distinguishing sarcastic statements from normal ones. An interesting task would be indeed to tackle this problem and see if sarcasm can really be detected or not.

C. SENTIMENT ANALYSIS: A TOOL OR A GOAL

When introduced in their respective papers, sarcasm detection is addressed by the authors as a mean to correct misclassified instances on a sentiment analysis exercise. However, in most of these works, sentiment analysis is used to actually extract features related to the polarity of the text. This leads to the obvious question: is sentiment analysis a tool, among others, to help identifying the sarcasm within a statement, or rather a target? In the latter case, where sarcasm is detected, the sentiment of the text is judged usually as the opposite of what it appears to be.

In a real use case, one would suggest that a data set is to be mined for sentiment analysis. Each piece of text is processed to identify its sentiment. Either sarcasm detection is first applied, and a set of texts judged as sarcastic are to be processed in a ''special'' way, whereas the rest is processed using the conventional sentiment analysis method,

[3] https://pan.webis.de/clef21/pan21-web/

TABLE 10. Sarcasm detection results reported in the literature on Twitter data sets.

Year	Authors	Accuracy	Precision	Recall	F1-score	AUC
2010	Davidov et al. [17]	0.947	0.912	0.756	0.827	
2011	González-Ibáñez et al. [108]	0.710				
2013	Lunando and Purwarianti [109]	0.541				
2013	Liebrecht et al. [45]					0.790
2013	Riloff et al. [110]		0.620	0.440	0.510	
2013	Reyes et al. [72]		0.780	0.740	0.760	
2014	Ptácek et al. [111]				0.947	
2014	Maynard and Greenwood [71]		0.910	0.910	0.910	
2014	Barbieri and Saggion [73]		0.760	0.760	0.760	
2014	Barbieri et al. [46]		0.750	0.760	0.760	
2014	Liu [112]					0.840
2014	Barbieri et al. [114]		0.900	0.900	0.900	
2015	Bamman and Smith [115]	0.851				
2015	Joshi et al. [19]		0.814	0.976	0.888	
2015	Rajadesingan et al. [52]	0.835				0.830
2015	Wang et al. [51]		0.537	0.704	0.603	
2015	Khattri et al. [62]		0.880	0.884	0.882	
2015	Fersini et al. [116]	0.942	0.782	0.899	0.836	
2015	Bharti et al. [18]		0.850	0.960	0.900	
2015	Bouazizi and Ohtsuki [9]	0.831	0.911	0.734		
2015	Muresan [50]	0.783			0.807	
2015	Ghosh et al. [117]		0.946	0.940	0.943	
2016	Bouazizi and Ohtsuki [47]	0.831	0.911	0.734		
2016	Ghosh and Veale [118]		0.919	0.923	0.921	
2015	Bouazizi and Ohtsuki [119]	0.831	0.911	0.734		
2016	Bharti et al. [75]		0.970	0.980	0.970	
2016	Zhang et al. [120]	0.907			0.907	
2016	Charalampakis et al. [121]		0.824	0.802	0.791	
2016	Farías et al. [122]				0.910	
2016	Abercrombie and Hovy [48]				0.710	0.640
2016	Poria et al. [53]				0.977	
2017	Al-Ghadhban et al. [125]		0.710	0.659	0.676	
2017	Dharwal et al. [126]				0.560	
2017	Gupta and Yang [127]		0.520	0.700	0.600	
2017	Mishra et al. [128]		0.765	0.753	0.757	
2017	Ghosh and Veale [129]		0.900	0.890	0.900	
2017	Mukherjee and Bala [132]	0.650			0.760	
2017	Jain et al. [133]	0.847	0.838	0.235		
2017	Prasad et al. [135]	0.818				
2017	Ghosh et al. [136]		0.773	0.765	0.763	
2018	Agrawal and An [84]				0.740	
2018	Ren et al. [85]				0.6328	
2018	Parmar et al. [86]		0.890	0.960	0.90	
2018	Samonte et al. [88]	0.936				
2018	A. Bohra et al. [89]				0.770	
2019	Liu et al. [90]		0.917	0.910	0.900	0.970
2019	Cai et al. [91]	0.834	0.766	0.842	0.802	
2019	Son et al. [92]	0.937	0.927	0.905	0.883	
2019	Tao et al. [93]	0.801	0.787	0.726	0.744	
2017	Bharti et al. [94]	0.870	0.848	0.836	0.842	
2020	Jain et al. [95]	0.927	0.895	0.907	0.891	
2020	Ren et al. [96]		0.874	0.893	0.872	
2021	H. Elgabry et al. [138]		0.686	0.670	0.676	
2021	Y. Zhang et al. [151]	0.708	0.708	0.708	0.708	
2021	D. Faraj et al. [139]	0.783	0.727	0.724	0.725	
2021	H. Nayel et al. [140]	0.746	0.705	0.560	0.594	
2021	M. Shrivastava et al. [141]	0.706	0.716	0.725	0.705	
2021	M. Bedi et al. [153]	0.873	0.865	0.636	0.711	
2021	A. Mahdaouy et al. [142]	0.768	0.727	0.712	0.718	
2021	C. Lou et al. [143]	0.903			0.819	
2021	A. Kumar et al. [154]	0.931	0.965	0.887	0.924	
2021	Y. Wu et al. [155]		0.752	0.746	0.745	
2021	F. Husain et al. [144]	0.773	0.713	0.681	0.692	
2021	M. S. Razali et al. [145]	0.940	0.950	0.940	0.940	
2021	F. Yao et al. [146]	0.832	0.770	0.825	0.797	
2021	X. Guo et al. [147]		0.367	0.465	0.410	
2022	A. Kamal et al. [148]	0.930	0.910	0.930	0.920	
2022	Y. Du et al. [149]	0.730			0.760	
2022	Z. Wen et al. [156]	0.761	0.761	0.759	0.759	

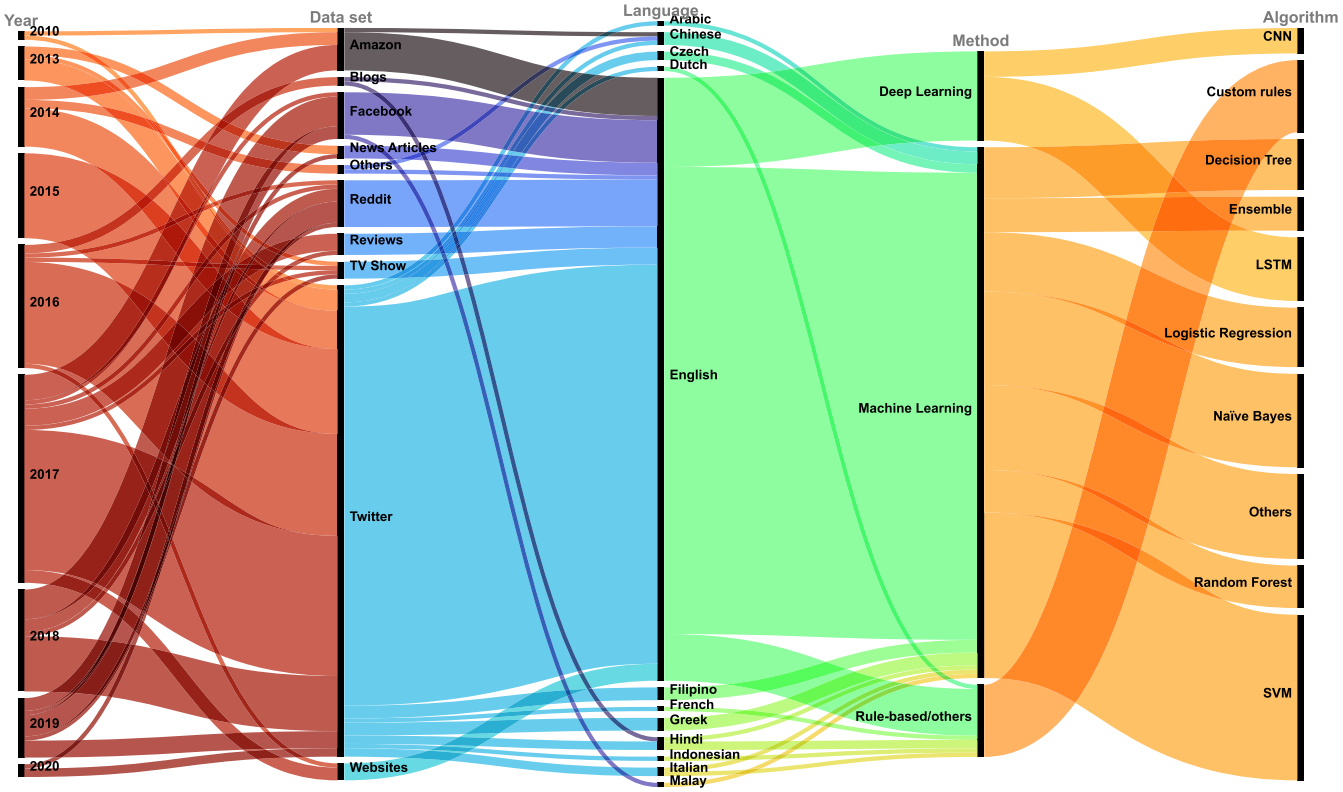


FIGURE 4. Distribution of the existing work over the years as a function of the data set used, language, method and algorithms used.

or sentiment analysis is applied regardless of whether sarcasm is present or not.

This brings the next question: how often sarcasm is used? In other words, is it really worthwhile to process tremendous amount of data for two tasks (sentiment analysis and sarcasm detection) if sarcasm is employed in a small fraction of the data. Other techniques such as sentiment quantification [36]–[39] were proposed in the literature to address partially wrong classified instances. The idea behind it is quite straightforward: when performed on a small data set, sentiment analysis will have a certain number of wrongly classified instances per class. Measuring how this error changes for different proportions of the classes in the data set could help learn how to rectify and interpret the results of classification when applied on a new data set. This could solve the problem of misclassification in a more global way, as usually identifying the polarity of the individual texts does not matter as much as identifying that of the entire data set.

D. SARCASM DETECTION AND EXPLAINABLE AI

EXplainable Artificial Intelligence (XAI), also known as interpretable AI, is a sub-field in Artificial Intelligence (AI) in which the AI models and/or results are presented in a way that can be understood by humans. Unlike the concept of a ''black box'' models in machine learning in which the human cannot, and sometimes is not even supposed to, know how the results are obtained and how the models are built. The concept of ''black box'' is not limited to machine learning model users, but also the designers themselves as, most of the time, they do not understand how the decisions made by their models is produced and how they can explain it to their model users. XAI addressed the idea of helping users understand how models work, by explaining the decision making process with reference to how a human would make the decision himself, and dismantling their misconceptions of how AI works. In theory, not only does this make the models more relevant but also builds a certain level of confidence towards them, allowing for a wider acceptance of AI among non-experts.

Generally speaking, different families of models have different levels of explainability as shown in FIGURE 5. Models such as decision tree ones are commonly known for being much more explainable than ones that rely on deep learning. On the other hand, as shown previously and as commonly agreed on, deep learning models are much more powerful in classification tasks than conventional machine learning ones.

With regards to sarcasm detection, a very important question rises when it comes to sarcasm detection: the intuition of the human brain develops to understand and recognize sarcasm is much more complicated than that developed to recognize sentiments for example. In the case of sentiment analysis, one would assume that the

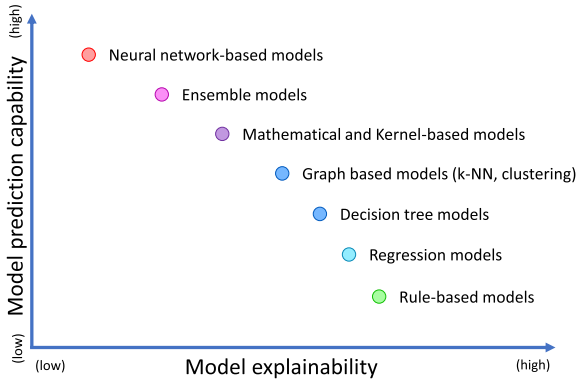


FIGURE 5. A representation of the relative explainability vs. prediction capability of the major machine learning model families.

presence of multiple positive words with no negation means the text containing them is mostly a positive one, and vice-versa. However, when it comes to sarcasm, the intuition is much more different, and generally speaking, the context needs to be taken into account to be able to detect it.

A few works, such as that of Riloff *et al.* [110] and Bouazizi and Ohtsuki [47] used an intuition similar to that used for sentiment analysis to build their models: if a positive word is co-existing in a text with a situation that is usually considered as negative, the text is likely to be sarcastic. The explainability of such models is, in that sense, much easier than that of transformers- or LSTM-based models.

That being said, tools for XAI have been receiving a great attention in the past few years. Namely, tools such as Shapley Additive exPlanations (SHAP) [183], [184] and Local Interpretable Model-agnostic Explanations (LIME) lime [185] have been widely accepted and used as tool for XAI. They way these models work, however, does not address the intuition behind feature engineering as much as they address the features themselves and their contribution to giving a good prediction, regardless or their meaning. They also operate at sample-level rather than model-level. In other words, the expected output of these tools is as follows: given an instance (e.g., tweet or text), how was the decision made to tell if it is sarcastic or not, and which features contributed to the identification of its class. With that in mind, an interesting task would be to see how much XAI could help narrow down the gap between how the model operates and how the human perceive, recognize and understand sarcasm, and whether or not these models follow similar intuitions like humans to tell if an instance is sarcastic or not.

A notable early attempt has been made by Kumar *et al.* [154], where the authors aimed to make the learning model used to detect sarcasm in conversations interpretable. However, in their work, the author relied mostly on individual words as indicator of sarcasm, which contradicts the intuition that sarcasm requires knowledge of relation between meanings in a sentence and the overall context for a better judgement as well as for a justified and convincing explanation.

E. LIMITS OF SARCASM DETECTION MODELS

In the previous sections, we tried to summarize the majority of the existing approaches and methods for sarcasm detection which we found. However, one question is yet to be answered: do these approaches and methods indeed perform as efficiently as they are expected when used on data collected outside of the context of their respective data sets? In other words, it is important to measure the generalizability of these approaches and whether they can indeed be used in real-world data sets other than the ones they were optimized for. For instance, given a model that has been developed using a data set collected a few years ago on a platform like Twitter, would it be able to detect sarcasm on nowadays tweets, or even statements collected elsewhere such as YouTube comments, Facebook posts or news websites? This leads the questions we previously raised in Section I:

- **[Q1]** Does the way people express sarcasm differ from one platform to another, and does it depend on the level of ''mastery'' of the language?
- **[Q2]** Does the way people express sarcasm evolve over time, in particular, on social media where sarcastic statements are ''driven'' by some influential users?
- **[Q3]** Is it safe to affirm that, for a given piece of text, if sarcasm is employed, the overall polarity of that text is the opposite of the apparent one?

With that in mind, in the next Section, we investigate in more detail this challenge, and aim to answer these questions with a set of experiments we run on data sets collected from different platforms and at different timestamps.

VI. EVOLUTION OF SARCASM OVER TIME AND ACROSS PLATFORMS

As stated above, in this section, we try to answer the following 3 questions, which we have previously introduced in Section I:

- **[Q1]** Does the way people express sarcasm differ from one platform to another, and does it depend on the level of ''mastery'' of the language?
- **[Q2]** Does the way people express sarcasm evolve over time, in particular, on social media where sarcastic statements are ''driven'' by some influential users?
- **[Q3]** Is it safe to affirm that, for a given piece of text, if sarcasm is employed, the overall polarity of that text is the opposite of the apparent one?

To do so, we will compare the results of some of the best-performing approaches on data sets collected from different sources at different time periods. We will start by introducing our experiment specifications and the data sets used. We then show the results of the different classification tasks that we run. Finally, we will discuss these results. Throughout the discussion, we will be answering the aforementioned questions.

A. EXPERIMENTAL SPECIFICATIONS

In the current work, we will use a deep learning and machine learning approaches to run the classification. We use

3 different implementations of sarcasm detection systems: A pattern-based approach [47] and two deep learning approaches that describe in more detail in this section. The first one uses [34] to identify sarcastic statements on 3 different data sets. However, the main focus of this work is not comparing the two approaches one to the other. On a first step we will try to identify whether sarcasm is expressed on different platforms the same way: a classification task will be run to try to guess for a given short text whether it is a sarcastic tweet, a sarcastic news headline or a sarcastic reddit comment. A similar classification task is performed to try to guess the temporal context of a given tweet: the classifier will try to guess whether a tweet was posted on the year 2015, 2017 or 2019. On a second step we use the pattern-based approach for sarcasm detection to identify the most common patterns used on each platform/temporal context, how often they are used on each platform, and whether there are patterns that are commonly used across the different platforms. Lastly, we take a closer look at the different data sets: we take a random set of samples from each and identify whether sarcastic tweets are polarity switcher or not: for each tweet, we check whether its actual sentiment polarity is the same as the one returned by a sentiment analysis tool (which does not recognize sarcasm) or not.

1) DATA

Two different data sets, manually annotated are used in this work:

- A set of tweets collected on three different points in time: mid 2015, mid 2017 and early 2019. This set will be referred to as **"Set I"**. These data have been collected using the Twitter streaming API by querying it for the hashtags *"#sarcasm"* and *"#not"*. In total, over 180,000 tweets were collected. The data have been manually checked and cleaned up by removing duplicates. They are also cleaned by removing all sorts of noise (e.g., non-English tweets, ones with images or URLs to external links). The resulting tweets are capped to a certain number to keep the data set balanced. These tweets, from each time span are split into two sub-sets as shown in TABLE 11: a training set and a test set. As can be seen, the data from 2017 are small in size compared to the other two time spans. This is because these data were collected over a smaller time span, thus resulting in a less quantity, even before cleaning.
- A set of sarcastic statements collected on three different platforms: Twitter, Reddit and some News websites. This set is referred to as **"Set II"**. The details of the different sub-sets of **"Set II"** are as follows:
  - The data for Twitter include for the most part tweets from the set previously describe (i.e., **"Set I"**).
  - Reddit data are available in several previous works, including IAC-SARC2 [158] and IAC-Subset [Walker *et al.*] [159].
  - Sarcastic news headlines are available online and a version of it can be obtained from Kaggle.[4]

  The overall data set has been cleaned as well. To make sure the classifier does not rely on features such as hashtags or the length of sentences, we made sure only textual components of the texts are kept and that statements from the different platforms have close average words per sentence. We have split this data set into two sub-sets as shown in TABLE 12: a training set and a test set.

TABLE 11. Sarcastic tweets from different time spans.

Class	Training set	Test set
Tweets from 2015	6000	4000
Tweets from 2017	3000	1631
Tweets from 2019	6000	4000
Total	**15 000**	**9 631**

TABLE 12. Sarcastic statements from different platforms.

Class	Training set	Test set
Twitter	8000	5355
Reddit	8000	2000
News headlines	8000	3724
Total	**24 000**	**11 079**

TABLE 13. Sarcastic and non-sarcastic tweets from different platforms.

Class	Sarcastic texts	Non-sarcastic texts
Twitter	10 000	10 000
Reddit	4 631	4 631
News headlines	10 000	10 000

As mentioned above, patterns are used, later in this paper, to identify whether there are commonly used expressions to express sarcasm on a given platform, and whether these platforms share common sarcastic patterns. Obviously, to generate patterns that are purely sarcastic, we need a set of non-sarcastic statements (for each type of data) against which we check the sarcastic ones. Therefore, we created 3 data sets (one for each platform: Twitter, Reddit and the news headlines in order) composed of sarcastic and non-sarcastic posts collected from the 3 platforms. These 3 data sets are referred to as **Sets "III-1"**, **"III-2"** and **"III-3"** respectively. The data set details are given in TABLE 13

2) HARDWARE AND SOFTWARE CONFIGURATIONS

For this experiment, we use a machine running on Windows 10 Pro, with the following hardware:

- **Processor:** Intel Core i7-6850K clocked at 3.6GHz
- **Memory:** 32GB RAM
- **Hard Drive:** 2TB HDD (7200rpm)
- **GPUs:** 2 × NVIDIA GeForce 1080 with 8GB of RAM each.

[4]https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection

The neural network is built using Python 3.7 and Pytorch. On the other hand, the patterns manually engineered are extracted using SENTA [35], an open source tool to extract features from texts. The tool mechanics have been modified to identify sarcastic texts from non-sarcastic ones instead of identifying the sentiment of the texts.

B. PROPOSED APPROACH

In Figure 6, we show the architecture of our proposed approach for sarcasm detection. The approach itself is an extension of a previous work of ours [47], in which we propose to use patterns for sarcasm detection. We will explain in detail what each part of the diagram means.

Given a piece of text t, we initially start by cleaning it (part **(a)** of the diagram). By cleaning, we refer to the process of removal of URLs and tags, replacing slang words and abbreviation by their corresponding full expression, etc. Two instances of the text are then created: the first one goes through the Neural Network (NN) shown in part **(b)** of the diagram while the second one is processed using SENTiment Analyzer (SENTA) [35] to extract pattern features as shown in part **(c)** of the diagram.

The upper part of the diagram, i.e., part **(b)**, corresponds to the neural network, through which goes the text and which identifies the class of the given text. The lower part, i.e., part **(c)**, corresponds to the ML part of the work, where engineered features (mainly patterns) are extracted.

The data sets **"I"** and **"II"** go through the neural networks, where the aim of the classification task, given a piece of text, is to identify whether it is a tweet, a reddit comment or a news headline (or at which time period it was posted), only using the textual information of the text.

The data sets **"III-1"**, **"III-2"** and **"III-3"** go through a different processing, where the aim is to extract the commonly used patterns to express sarcasm on each platform.

In the rest of this section, we describe in more details each of the steps.

1) PRE-PROCESSING

The pre-processing phase consists of several basic tasks to clean up the data sets used for training and testing. They include:

- Removing all the URLs, tags and all non-textual components,
- Replacing slang words and abbreviation by their corresponding English words and expressions,
- Fixing all detected typos and excessive punctuation marks usages.

An additional pre-processing step is made for texts that will be used for pattern extraction: all punctuation marks are removed and names are replaced by a simple expression to refer to them. In addition, as mentioned above, all the sets, whether they are used for training and for test are pre-processed the same way.

2) NEURAL NETWORK

In this part of the work, we use the pre-trained language model implemented by Howard and Ruder [186], which we then fine-tune for our current task. The original model was trained on the WikiText-103 data set [187]. This corpus is composed of 28 595 pre-processed Wikipedia articles and contains 103 million distinct tokens (words). Howard and Ruder [186] implemented an AWD-LSTM (Averaged Stochastic Gradient Descent Weight-Dropped LSTM), a regular LSTM with various tuned dropout hyperparameters [187]. Figure 7 shows the structure of the AWD-LSTM. The model is composed of an embedding layer, followed by 3 stacked LSTM layers and a softmax layer. The embedding size is 400 and each LSTM layer has 1152 activations.

As shown in the figure, the first step is to load the model as it is. This step is done by simply calling the pre-trained model. The model is downloaded alongside with its training weights.

In the second step, we start the fine-tuning of the language model, by continuing the training on twitter+reddit+ news headline-like data. To do so, we used our whole data sets, alongside with more data collected from Twitter and Reddit, with no label. This goal of this step is to make the language model learn the specific features of the language used in these platforms so that it could recognize how sentences are structured and learn newer words such as slangs.

In the final step, the language model is adjusted for classification. From the language model, the softmax layer is cut off, and a linear block whose activation is set to softmax is added after the 3 LSTM layers. We fine-tune the model using the gradual unfreezing, discriminative learning rates, and the slanted triangular learning rate.

3) PATTERN EXTRACTION

Patterns are extracted with SENTA [35] as we have described previously. We have referred to the work of Bouazizi and Ohtsuki [47] to extract them.

Pattern features identify and quantify the full expressions that are commonly used to express sarcasm. A pattern is defined as a generic sequence of words or expressions. They are collected according to specific rules as described in [47]: all words are divided into 2 groups:

- *"CI"*: this group contains the words whose content is important, and
- *"GFI"*: this group contains the words whose grammatical function is important.

Words whose grammatical function is important are replaced by some expression [47], whereas words whose content is important are kept as they are. The classification of words into one of these categories is based on their Part-of-Speech (PoS) tag. A pattern p obey the rule that their length should be within a certain length range [47]:

$$L_{Min} \leq \text{Length}(p) \leq L_{Max} \tag{2}$$

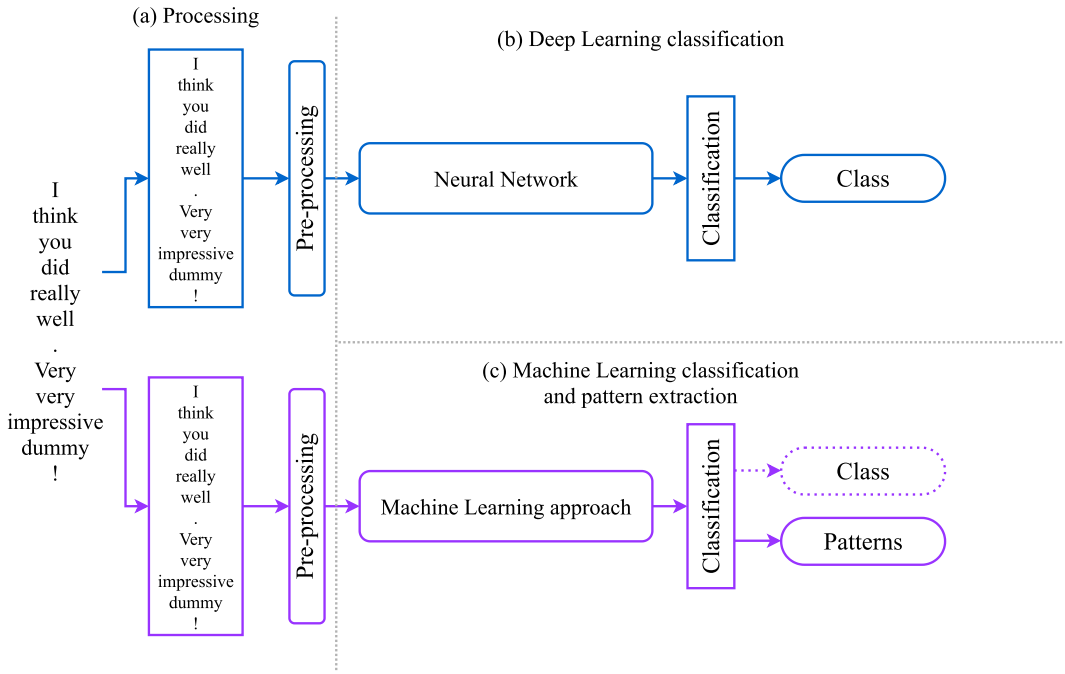


FIGURE 6. Architecture of the proposed approach. The part (a) refers to the pre-processing part. The part (b) refers to the neural network architecture and the classification. The part (c) refers to the pattern extraction and comparison.

where L_{Min} and L_{Max} are the minimum and maximum length of a pattern in terms of words. Patterns are extracted from the sarcastic texts, filtered according to certain rules: a pattern p must occur N_{occ} times, and should not appear, not even once, on a non-sarcastic text. Throughout our experiments, we use the same values for parameters L_{Min}, L_{Max} and N_{occ} as in [47].

An example of a pattern extracted from the following sentence *"I love it when people take me for granted"* would be *[PRONOUN love PRONOUN when NOUN VERB]*.

Patterns extracted from the different platforms are compared to each other, to see if there are some commonly used patterns on all the platforms, or whether a specific platform has a different behavior when compared to others (**"Set I"**). Similarly, patterns extracted at different points in time from twitter are compared to each other for the same purpose (**"Set II"**).

C. EXPERIMENTAL RESULTS

In the first set of experiments, we run a classification task using **"Set I"** and **"Set II"**. The aim is to identify where (or when) a given piece of text was posted. This is of a great importance to highlight later on to what extent the use of patterns would be useful, and whether or not approaches for sarcasm detection on Twitter, the most studied platform in this context, can be applied on other ones.

To evaluate the performance of classification, we use the following KPIs:

- TPR,
- Precision,
- Recall, and
- F-Score.

which we have previously defined in Section IV-D.

These KPIs are measured at class-level as well as on the test sets in their entirety.

TABLE 14. Classification accuracy, precision, recall and f-measure for different platforms.

	TPR	Precision	Recall	F-Score	ROC AUC
Twitter	76.43%	89.31%	76.43%	82.37%	72.34%
Reddit	75.00%	55.33%	75.00%	63.68%	77.83%
News headlines	98.25%	96.67%	98.25%	97.46%	84.30%
Overall	**83.51%**	**85.65%**	**83.51%**	**84.57%**	**77.35%**

TABLE 15. Confusion matrix of the classification for different platforms.

Class	Classified as		
	Twitter	Reddit	News headlines
Twitter	**4093**	1181	81
Reddit	455	**1500**	45
News headlines	35	30	**3659**

1) SARCASM ACROSS PLATFORMS

We initially run the classification of texts from different platforms agains each other. The classification is done using the DL part of Figure 6. The classification TPR, precision, recall and F1-measure are given in TABLE 14. The confusion matrix of classification is given in TABLE 15.

As we can observe, it is relatively easy to distinguish sarcastic statements from different platforms from each other. To recall, this is not due to the use of non-textual components

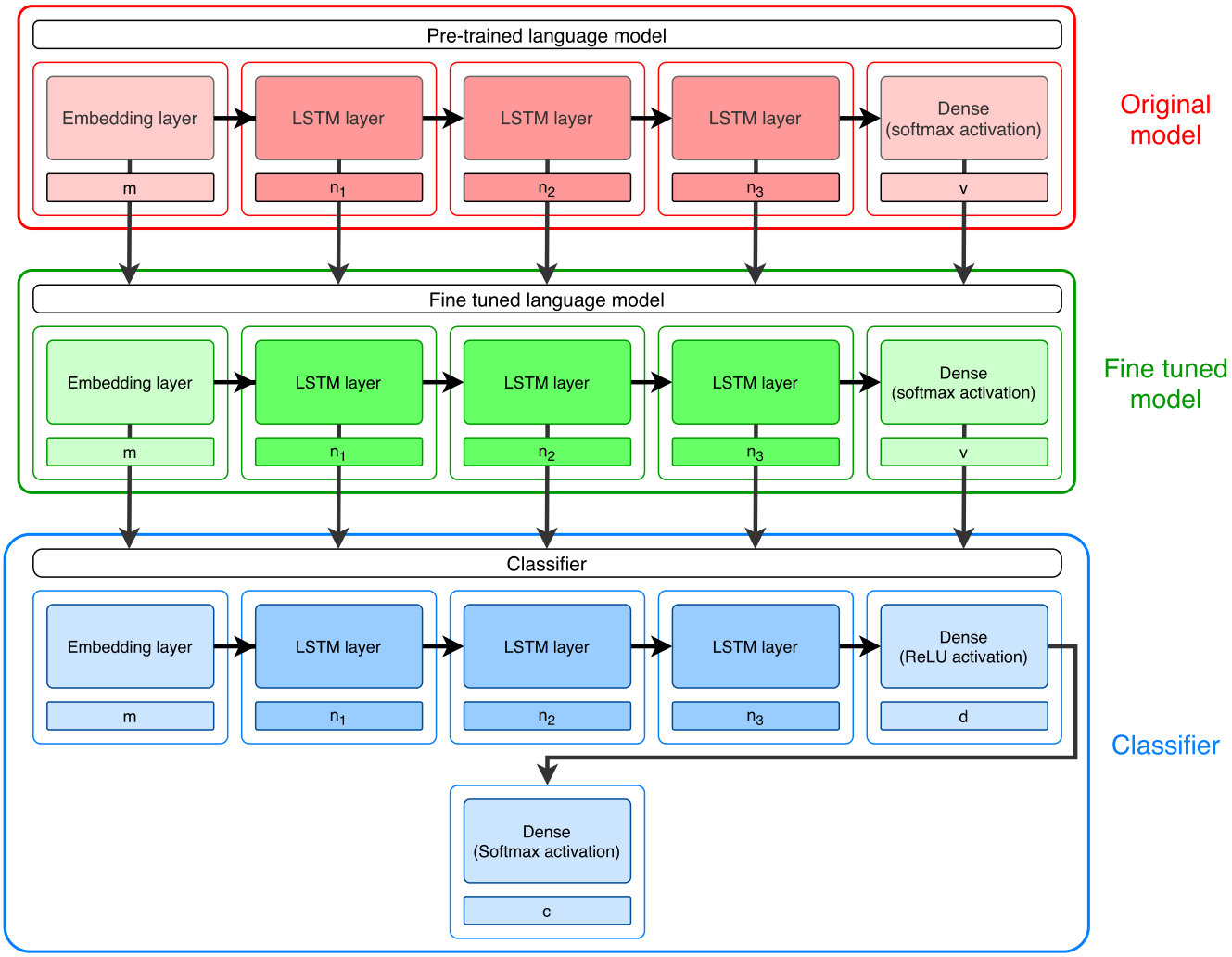


FIGURE 7. Architecture of the neural network used and the 3 steps to fine tune it to our current task: 1) load the original language model, 2) fine tune it on corpus made of data similar to ours, 3) add a softmax layer, and train it to perform a classification task.

or punctuation because these have been removed prior to the classification. This is not due to the difference in length of the statements as they are of similar average lengths. One reason for the distinction between sarcastic news headlines and sarcastic tweets and reddit comments is how formal the language employed is. This explains why the sarcastic news headlines have such high precision and recall levels (96.67% and 98.25% respectively). However, being both causal user-generated, tweets and reddit comments share a lot in common, yet are distinguishable from each other to a certain level. Later in this work, we will discuss this particular point.

2) SARCASM OVER TIME

A more interesting task would be to identify sarcastic statements at different points in time. While such task can be highly biased due to the trending topics at these different points in time, we still believe that the way sarcasm is expressed changes, or rather evolves over the time.

TABLE 16. Classification accuracy, precision, recall and f-measure for different time points.

	TPR	Precision	Recall	F-Score	ROC AUC
Tweets from 2015	70.98%	68.77%	70.98%	69.86%	73.12%
Tweets from 2017	44.70%	48.70%	44.70%	46.61%	56.40%
Tweets from 2019	68.48%	68.37%	68.48%	68.42%	67.41%
Overall	**65.49%**	**65.21%**	**65.49%**	**65.35%**	**68.18%**

TABLE 17. Confusion matrix of the classification for different time points.

Class	Classified as		
	T from 2015	T from 2017	T from 2019
T from 2015	**2839**	352	809
T from 2017	444	**729**	458
T from 2019	845	416	**2739**

The classification performance is shown in TABLE 16 and its confusion matrix is given in TABLE 17.

The classification results show that it is possible to distinguish between sarcastic tweets posted at different points in time. As we said above, this might be partially biased by the

TABLE 18. Classification accuracy, precision, recall and f-score of the models $\mathbf{M_T^{2015}}$, $\mathbf{M_T^{2017}}$ and $\mathbf{M_T^{2019}}$ on the different test sets.

	Accuracy	Precision	Recall	F-Score	ROC AUC
Using the model $\mathbf{M_T^{2015}}$					
Tweets from 2015	89.20%	91.70%	86.20%	88.87%	87.24%
Tweets from 2017	74.06%	69.04%	74.34%	71.59%	81.01%
Tweets from 2019	69.73%	69.40%	70.55%	69.97%	74.21%
Using the model $\mathbf{M_T^{2017}}$					
Tweets from 2015	74.06%	69.04%	75.55%	76.51%	76.14%
Tweets from 2017	86.27%	84.67%	83.96%	84.31%	83.94%
Tweets from 2019	73.00%	72.14%	74.95%	73.52%	75.55%
Using the model $\mathbf{M_T^{2019}}$					
Tweets from 2015	72.35%	72.69%	71.60%	72.14%	73.11%
Tweets from 2017	78.85%	71.83%	85.36%	78.01%	77.81%
Tweets from 2019	80.58%	83.95%	75.60%	79.56%	80.11%

topic of the tweet itself. However, later on, we show that the patterns commonly used to express sarcasm change. In other words, people ''learn'' from influential users, who come up with new trends of how to express sarcasm.

3) MODEL GENERALIZATION

Given the results obtained above, our next target is to identify whether sarcasm detection models are generalizable. In other word, given a model trained and optimized on one data set, we want to see how good it is when evaluated on another data set. With the major differences between the different platforms, we limit our study to the same platform, and evaluate the models trained on a data set from one time period in identifying sarcasm on data sets from the other two time periods.

We trained 3 versions of neural networks shown in Figure 7 on Twitter data sets from **Set I**. We refer to the three models trained on tweets from 2015, 2017 and 2019 as $\mathbf{M_T^{2015}}$, $\mathbf{M_T^{2017}}$ and $\mathbf{M_T^{2019}}$, respectively. Each model has been training on its corresponding training set, and validated on its corresponding test set. In Table 18, we report the results of classification using each of the models on the 3 different test sets (including the one from its time span). Note that the precision, recall and F-score are reported for the class ''sarcastic''. As can be seen, each model performs best when used for the data collected from its time span. For instance, the accuracy of the model $\mathbf{M_T^{2015}}$ on data collected in 2015 reached 89.20%. This accuracy drops significantly when the model is evaluated on data collected in 2017 and 2019, reaching 74.06% and 69.73%, respectively. This behavior is observed also when using the two other models (i.e., $\mathbf{M_T^{2017}}$ and $\mathbf{M_T^{2019}}$).

To answer our question about the generalizabily of the models, it is fair to conclude from the results obtained in Table 18 that a model trained on data from a certain time span could be indeed used to classify data from another time span. However, one would assume that the expected performance is far lower than that obtained during the training. Later, in subsection VI-C5, we address the main reasons for such drop in performance.

4) SARCASTIC PATTERNS

In this sub-section, we explore the idea that suggests that people use similar phrase and sentences to express sarcasm.

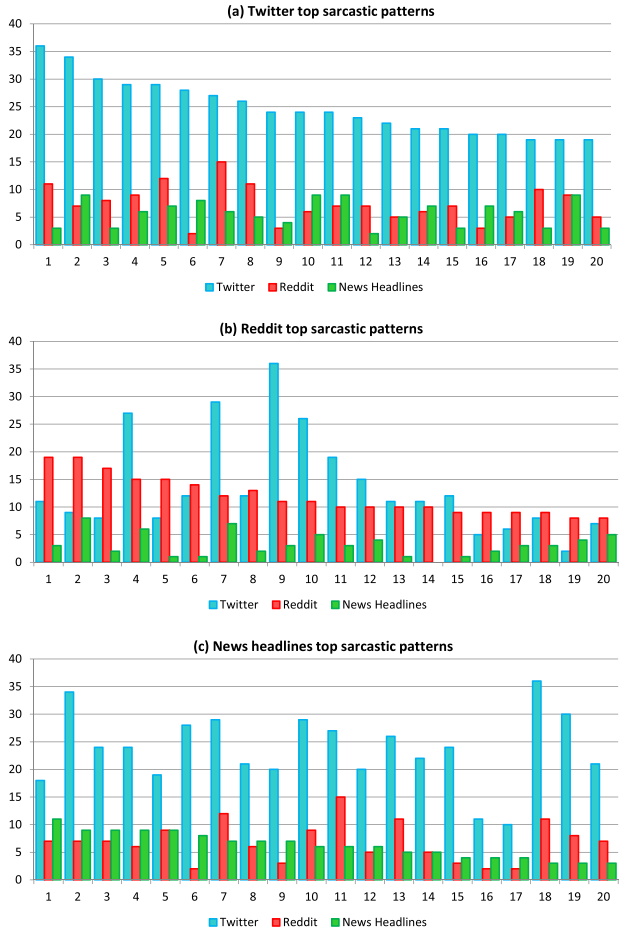
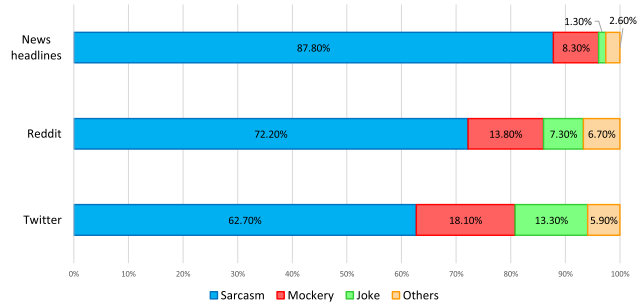


FIGURE 8. Common patterns ordered by their usage on each platform.

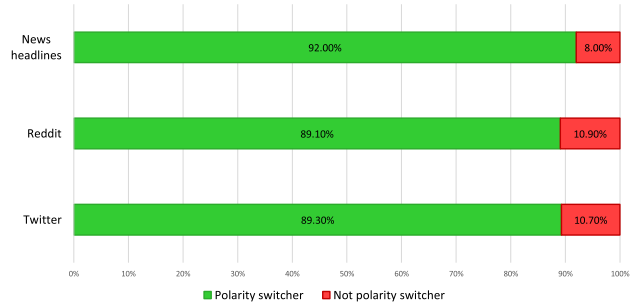
This is the key point behind several works on the detection of sarcasm [17], [43], [47]. In their respective works, they used the term ''pattern'' to refer to a generalization of these expressions used to express sarcasm. As described previously in Section VI-B3, patterns are collected by transforming the texts in a way that abstract them from their context, while keeping the overall grammatical construction of the sentences. Here we used the **Sets "III-1"**, **"III-2"** and **"III-3"** to collect the most common sarcastic patterns in each platform of our data set (i.e., Twitter, Reddit and the news headlines).

In Figure 8, we show the top 20 patterns used in each platform ordered by their occurrence number in each platform, as well as their occurrence number in the other platforms. Figure 8-(a) shows patterns that are mostly used in Twitter, Figure 8-(b) shows those that are mostly used in Reddit, and Figure 8-(c) shows those that are mostly used in news headlines.

As we can observe, patterns extracted from tweets are the most abundant ones. By far, they outnumber those extracted from the other platforms. This goes along with our intuition early that suggests that sarcastic expressions are mostly learned and re-used. This also explains how approaches that



(a) Intent of the use of sarcasm



(b) Effect of sarcasm on the sentiment polarity of the text

FIGURE 9. An in-depth look at a sample of texts manually labeled as "Sarcastic."

rely on patterns to detect sarcasm are very good at detecting sarcasm in Twitter, but present lower performance on other platforms [47].

On the other hand, it is interesting to note that the top patterns are almost all common to all the platforms, meaning that these expressions are used in all the platforms despite how formal or informal the language used is. This means that some patterns are ''universally'' agreed-on as sarcastic.

5) SARCASM: IS IT REALLY SARCASM?

As we have explained early on in this paper, sarcasm is a sophisticated form of speech which requires a higher-than-average intelligence to make and to understand [15]. Several suggestions have been made towards whether sarcasm, the way it is expressed in social media is the ultimate form of sarcasm [44], [47]. Early in this paper, we introduced **[Q1]** questioning as well whether sarcasm requires a certain mastery of the language for it to be employed correctly.

As we have seen in the previous sub-section, users of Twitter tend to overuse certain expressions, making them more of abused clichés and less of contemptuous phrases.

Here, we collected some random samples of texts from our data set, and went through them to identify whether or not they really are sarcastic. In addition, being one of the reasons sarcasm has been studied in the first place, we also checked, for each of the texts, whether it is a polarity switcher or not, answering the question **[Q3]**.

In Figure 9, we classify the studied samples of texts from into 1 of 4 of classes, identifying whether or not what the annotators labeled as sarcastic is indeed so. In the upper part of the figure (Figure 9-a), we show the proportion of tweets, reddit posts and news headlines annotated as sarcastic which have been recognized by more rigorous annotators as really being sarcastic. Note that the data sets were partially collected from various sources found online and partially annotated through the services of CrowdFlower.[5] As shown, a little less than 63% of these tweets are actually sarcastic. The remaining ones are most directly mocking (a person usually) or jokes. Similarly, 72.2% of the Reddit comments annotated as sarcastic were indeed identified sarcastic after rigorous check, and 87.8% of news headlines annotated as sarcastic were identified as so.

On the other hand, as shown in Figure 9-b, sarcasm is indeed a polarity switcher for 89.3% of the tweets where it is employed. In other words, the actual polarity of the tweet is the opposite of that returned by a sentiment analyzer. Sarcasm has also been identified as a polarity switcher on 89.1% and 92.0%, respectively in Reddit and the news headlines, respectively. It is no surprise that news headlines were ones that had more accurately sarcastic statements, and that sarcasm is more of a polarity switcher when compared to the other two platforms. This has been targeted in the previous sub-section.

To recapitulate, the answers for the questions we have investigated could be as follows:

- **[A1]** The way sarcasm is expressed indeed differs from one platform to another. This is due to several reasons which include, but are not limited to, the mastery of language, how influenced users are by others, etc.
- **[A2]** The way sarcasm is expressed in Twitter does change over the time. More interestingly, Twitter is indeed the platform where pattern-based approaches for sarcasm detection are the most effective.
- **[A3]** Within the context of the data sets we have explored and used in this work, it is safe to affirm that sarcasm is a polarity switcher with a very high probability.

VII. CONCLUSION

In this paper, we have investigated the topic of sarcasm detection on different platforms and over time. We have studied the different ways sarcasm is expressed on 3 different platforms: Twitter, Reddit and news headlines. Our experiments show that sarcasm is indeed expressed differently on these platforms. They have also shown that the way it is expressed at separate periods of time is different. Finally we have explored the idea of using sarcasm as a polarity switcher, and confirmed that sarcasm can be used as a polarity switcher if detected.

REFERENCES

[1] S. Homoceanu, M. Loster, C. Lofi, and W.-T. Balke, "Will I like it? Providing product overviews based on opinion excerpts," in *Proc. IEEE 13th Conf. Commerce Enterprise Comput. (CEC)*, Sep. 2011, pp. 26–33.

[5] https://www.crowdflower.com/

[2] U. R. Hodeghatta, "Sentiment analysis of Hollywood movies on Twitter," in *Proc. IEEE/ACM ASONAM*, Aug. 2013, pp. 1401–1404.
[3] H. A. Aldahawi and S. M. Allen, "Twitter mining in the oil business: A sentiment analysis approach," in *Proc. Int. Conf. Cloud Green Comput.*, Sep. 2013, pp. 581–586.
[4] E. Ferrara, H. Chang, E. Chen, G. Muric, and J. Patel, "Characterizing social media manipulation in the 2020 U.S. Presidential election," *1st Monday*, vol. 25, no. 11, Oct. 2020. [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/11431/9993
[5] G. Barkur and G. B. Kamath, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India," *Asian J. Psychiatry*, vol. 51, Jun. 2020, Art. no. 102089.
[6] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074–181090, 2020.
[7] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114155.
[8] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Gener. Comput. Syst.*, vol. 106, pp. 92–104, May 2020.
[9] M. Bouazizi and T. Ohtsuki, "Opinion mining in Twitter how to make use of sarcasm to enhance sentiment analysis," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2015, pp. 1594–1597.
[10] M. S. Razali, A. A. Halin, N. M. Norowi, and S. C. Doraisamy, "The importance of multimodality in sarcasm detection for sentiment analysis," in *Proc. IEEE 15th Student Conf. Res. Develop. (SCOReD)*, Dec. 2017, pp. 56–60.
[11] R. Giora, "On irony and negation," *Discourse Process.*, vol. 19, no. 2, pp. 239–264, 1995.
[12] E. Camp, "Sarcasm, pretense, and the semantics/pragmatics distinction*," *Noûs*, vol. 46, no. 4, pp. 587–634, Dec. 2012.
[13] K. Durkin, *Developmental Social Psychology: From Infancy to Old Age*. Hoboken, NJ, USA: Blackwell, 1995.
[14] M. L. Sundberg, J. Michael, J. W. Partington, and C. A. Sundberg, "The role of automatic reinforcement in early language acquisition," *Anal. Verbal Behav.*, vol. 13, no. 1, pp. 21–37, Apr. 1996, doi: 10.1007/BF03392904.
[15] L. Huang, F. Gino, and A. D. Galinsky, "The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients," *Organizational Behav. Hum. Decis. Processes*, vol. 131, pp. 162–177, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S074959781500076X
[16] G. Deliens, K. Antoniou, E. Clin, and M. Kissine, "Perspective-taking and frugal strategies: Evidence from sarcasm detection," *J. Pragmatics*, vol. 119, pp. 33–45, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037821661730022X
[17] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcasm in Twitter and Amazon," in *Proc. 14th Conf. Comput. Natural Lang. Learn.*, New York, NY, USA, 2010, pp. 107–116.
[18] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in Twitter data," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2015, pp. 1373–1380.
[19] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, vol. 2, 2015, pp. 757–762. [Online]. Available: https://www.aclweb.org/anthology/P15-2124
[20] M. V. Mantyla, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Comput. Sci. Rev.*, vol. 27, pp. 16–32, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574013717300606
[21] J. A. Richmond, "Spies in ancient Greece," *Greece Rome*, vol. 45, no. 1, pp. 1–18, Apr. 1998. [Online]. Available: http://www.jstor.org/stable/643204
[22] D. D. Droba, "Methods used for measuring public opinion," *Amer. J. Sociol.*, vol. 37, no. 3, pp. 410–423, Nov. 1931, doi: 10.1086/215733.
[23] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Jul. 2002, pp. 79–86. [Online]. Available: https://www.aclweb.org/anthology/W02-1011
[24] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. Int. AAAI Conf. Weblogs Social Media*, Washington, DC, USA, 2010, p. 559. [Online]. Available: http://www.scribd.com/doc/31302916/From-Tweets-to-Polls-Linking-Text-Sentiment-to-Public-Opinion-Time-Series
[25] M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons," in *Proc. IISA, 5th Int. Conf. Inf., Intell., Syst. Appl.*, Jul. 2014, pp. 94–97.
[26] L. Martin-Domingo, J. C. Martín, and G. Mandsberg, "Social media as a resource for sentiment analysis of airport service quality (ASQ)," *J. Air Transp. Manage.*, vol. 78, pp. 106–115, Jul. 2019.
[27] W. Duan, Q. Cao, Y. Yu, and S. Levy, "Mining online user-generated content: Using sentiment analysis technique to study hotel service quality," in *Proc. 46th Hawaii Int. Conf. Syst. Sci.*, Jan. 2013, pp. 3119–3128.
[28] N. A. Vidya, M. I. Fanany, and I. Budi, "Twitter sentiment to analyze net brand reputation of mobile phone providers," *Proc. Comput. Sci.*, vol. 72, pp. 519–526, Jan. 2015.
[29] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu, "Identifying breakpoints in public opinion," in *Proc. 1st Workshop Social Media Anal. (SOMA)*, New York, NY, USA, 2010, pp. 62–66, doi: 10.1145/1964858.1964867.
[30] M. Boia, B. Faltings, C.-C. Musat, and P. Pu, "A :) is worth a thousand words: How people attach sentiment to emoticons and words in tweets," in *Proc. Int. Conf. Social Comput.*, Sep. 2013, pp. 345–350.
[31] K. Manuel, K. V. Indukuri, and P. R. Krishna, "Analyzing internet slang for sentiment mining," in *Proc. 2nd Vaagdevi Int. Conf. Inf. Technol. Real World Problems*, Dec. 2010, pp. 9–11.
[32] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "What emotions do news articles trigger in their readers?" in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 2007, pp. 733–734, doi: 10.1145/1277741.1277882.
[33] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "Emotion classification of online news articles from the Reader's perspective," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Dec. 2008, pp. 220–226.
[34] M. Bouazizi and T. Ohtsuki, "Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.
[35] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multi-class sentiment analysis in Twitter," *IEEE Access*, vol. 5, pp. 20617–20639, 2017.
[36] J. Barranquero, J. Díez, and J. José del Coz, "Quantification-oriented learning based on reliable classifiers," *Pattern Recognit.*, vol. 48, no. 2, pp. 591–604, Feb. 2015.
[37] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramirez-Quintana, "Quantification via probability estimators," in *Proc. IEEE Int. Conf. Data Mining*, 2010, pp. 737–742.
[38] A. Esuli and F. Sebastiani, "Optimizing text quantifiers for multivariate loss functions," *ACM Trans. Knowl. Discovery Data*, vol. 9, no. 4, pp. 1–27, Jun. 2015.
[39] G. Forman, "Quantifying counts and costs via classification," *Data Mining Knowl. Discovery*, vol. 17, no. 2, pp. 164–206, Oct. 2008.
[40] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis in Twitter: What if classification is not the answer," *IEEE Access*, vol. 6, pp. 64486–64502, 2018.
[41] J. Tepperman, D. Traum, and S. S. Narayanan, "'Yeah right': Sarcasm recognition for spoken dialogue systems," in *Proc. INTERSPEECH*, Sep. 2006, pp. 1838–1841.
[42] R. J. Kreuz and G. M. Caucci, "Lexical influences on the perception of sarcasm," in *Proc. Workshop Comput. Approaches Figurative Lang. (FigLanguages)*, New York, NY, USA, 2007, pp. 1–4.
[43] O. Tsur, D. Davidov, and A. Rappoport, "CWSM—A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews," in *Proc. ICWSM*, Washington, DC, USA, 2010, pp. 162–169.
[44] S. Lukin and M. Walker, "Really? Well. Apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue," in *Proc. Workshop Lang. Anal. Social Media*. Atlanta, GA, USA, Jun. 2013, pp. 30–40.

[45] C. Liebrecht, F. Kunneman, and A. van den Bosch, "The perfect solution for detecting sarcasm in tweets #not," in *Proc. 4th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Atlanta, GA, USA, Jun. 2013, pp. 29–37.
[46] F. Barbieri, F. Ronzano, and H. Saggion, "Italian irony detection in Twitter: A first approach," in *Proc. 1st Italian Conf. Comput. Linguistics (CLiC-It)*. Paris, France: Pisa Univ. Press, Dec. 2014, pp. 28–32.
[47] M. Bouazizi and T. O. Ohtsuki, "A pattern-based approach for sarcasm detection on Twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016.
[48] G. Abercrombie and D. Hovy, "Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations," in *Proc. ACL Student Res. Workshop*, Berlin, Germany, 2016, pp. 107–113. [Online]. Available: https://www.aclweb.org/anthology/P16-3016
[49] A. Joshi, V. Tripathi, P. Bhattacharyya, and M. J. Carman, "Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends,'" in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, Berlin, Germany, 2016, pp. 146–155. [Online]. Available: https://www.aclweb.org/anthology/K16-1015
[50] S. Muresan, R. Gonzalez-Ibanez, D. Ghosh, and N. Wacholder, "Identification of nonliteral language in social media: A case study on sarcasm," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 11, pp. 2725–2737, 2016. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23624
[51] Z. Wang, Z. Wu, R. Wang, and Y. Ren, "Twitter sarcasm detection exploiting a context-based model," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Cham, Switzerland: Springer, 2015, pp. 77–91. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-26190-4_6#citeas
[52] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on Twitter: A behavioral modeling approach," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, New York, NY, USA, Feb. 2015, pp. 97–106, doi: 10.1145/2684822.2685316.
[53] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," 2016, *arXiv:1610.08815*.
[54] B. C. Wallace, D. K. Choe, and E. Charniak, "Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, vol. 1, 2015, pp. 1035–1044. [Online]. Available: https://www.aclweb.org/anthology/P15-1100
[55] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? Using satirical cues to detect potentially misleading news," in *Proc. 2nd Workshop Comput. Approaches Deception Detection*, San Diego, CA, USA, Jun. 2016, pp. 7–17. [Online]. Available: https://www.aclweb.org/anthology/W16-0802
[56] C. L. Davis, K. Oishi, A. V. Faria, J. Hsu, Y. Gomez, S. Mori, and A. E. Hillis, "White matter tracts critical for recognition of sarcasm," *Neurocase*, vol. 22, no. 1, pp. 22–29, Jan. 2016.
[57] C. Lomen-Hoerth, T. Anderson, and B. Miller, "The overlap of amyotrophic lateral sclerosis and frontotemporal dementia," *Neurology*, vol. 59, no. 7, pp. 1077–1079, Oct. 2002.
[58] M. Staios, F. Fisher, A. K. Lindell, B. Ong, J. Howe, and K. Reardon, "Exploring sarcasm detection in amyotrophic lateral sclerosis using ecologically valid measures," *Frontiers Hum. Neurosci.*, vol. 7, p. 178, Jan. 2013.
[59] J. Jorgensen, "The functions of sarcastic irony in speech," *J. Pragmatics*, vol. 26, no. 5, pp. 613–634, Nov. 1996.
[60] M. A. Seckman and C. J. Couch, "Jocularity, sarcasm, and relationships: An empirical study," *J. Contemp. Ethnography*, vol. 18, no. 3, pp. 327–344, 1989.
[61] D. C. Littman and J. L. Mey, "The nature of irony: Toward a computational model of irony," *J. Pragmatics*, vol. 15, no. 2, pp. 131–151, Feb. 1991.
[62] A. Khattri, A. Joshi, P. Bhattacharyya, and M. Carman, "Your sentiment precedes you: Using an author's historical tweets to predict sarcasm," in *Proc. 6th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2015, pp. 25–30.
[63] P. Brown, S. C. Levinson, and S. C. Levinson, *Politeness: Some Universals in Language Usage*, vol. 4. Cambridge, U.K.: Cambridge Univ. Press, 1987.
[64] J. Jorgensen, G. A. Miller, and D. Sperber, "Test of the mention theory of irony," *J. Exp. Psychology: Gen.*, vol. 113, no. 1, p. 112, 1984.
[65] D. Sperber and D. Wilson, "Irony and the use-mention distinction," *Philosophy*, vol. 3, pp. 143–184, Aug. 1981.
[66] J. Gumperz, "The linguistic bases of communicative competence," in *Analyzing Discourse: Text and Talk*. 1982, pp. 323–334.
[67] R. J. Kreuz and S. Glucksberg, "How to be sarcastic: The echoic reminder theory of verbal irony," *J. Exp. Psychol., Gen.*, vol. 118, no. 4, p. 374, 1989.
[68] K. P. Rankin, A. Salazar, M. L. Gorno-Tempini, M. Sollberger, S. M. Wilson, D. Pavlic, C. M. Stanley, S. Glenn, M. W. Weiner, and B. L. Miller, "Detecting sarcasm from paralinguistic cues: Anatomic and cognitive correlates in neurodegenerative disease," *NeuroImage*, vol. 47, no. 4, pp. 2005–2015, Oct. 2009.
[69] R. W. Gibbs, "On the psycholinguistics of sarcasm," *J. Exp. Psychol., Gen.*, vol. 115, no. 1, p. 3, 1986.
[70] S. McDonald, "Exploring the process of inference generation in sarcasm: A review of normal and clinical studies," *Brain Lang.*, vol. 68, no. 3, pp. 486–506, Jul. 1999.
[71] D. Maynard and M. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 4238–4243.
[72] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in Twitter," *Lang. Resour. Eval.*, vol. 47, no. 1, pp. 239–268, Mar. 2013.
[73] F. Barbieri and H. Saggion, "Modelling irony in Twitter: Feature analysis and evaluation," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 4258–4264.
[74] A. Reyes and P. Rosso, "On the difficulty of automatically detecting irony: Beyond a simple case of negation," *Knowl. Inf. Syst.*, vol. 40, no. 3, pp. 595–614, 2014.
[75] S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection in tweets streamed in real time: A big data approach," *Digit. Commun. Netw.*, vol. 2, no. 3, pp. 108–121, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S235286481630027X
[76] P. Rockwell, *The Effects of Cognitive Complexity and Communication Apprehension on the Expression and Recognition of Sarcasm*. Hauppauge, NY, USA: Nova Science Publishers, 2007.
[77] M. Basavanna, *Dictionary of Psychology*. New Delhi, India: Allied Publishers, 2000.
[78] M. Toplak and A. N. Katz, "On the uses of sarcastic irony," *J. Pragmatics*, vol. 32, no. 10, pp. 1467–1488, Sep. 2000.
[79] H. S. Cheang and M. D. Pell, "Recognizing sarcasm without language: A cross-linguistic study of English and Cantonese," *Pragmatics Cognition*, vol. 19, no. 2, pp. 203–223, 2011.
[80] P. Rockwell and E. M. Theriot, "Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis," *Commun. Res. Rep.*, vol. 18, no. 1, pp. 44–52, 2001.
[81] E. Filatova, "Irony and sarcasm: Corpus generation and analysis using crowdsourcing," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 392–398.
[82] E. Sulis, D. I. H. Faras, P. Rosso, V. Patti, and G. Ruffo, "Figurative messages and affect in twitter: Differences between irony, sarcasm and not," *Knowl.-Based Syst.*, vol. 108, pp. 132–143, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705116301320
[83] J. Karoui, B. Farah, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles, "Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Valencia, Spain, vol. 1, 2017, pp. 262–272. [Online]. Available: https://www.aclweb.org/anthology/E17-1025
[84] A. Agrawal and A. An, "Affective representations for sarcasm detection," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, Jun. 2018, pp. 1029–1032.
[85] Y. Ren, D. Ji, and H. Ren, "Context-augmented convolutional neural networks for Twitter sarcasm detection," *Neurocomputing*, vol. 308, pp. 1–7, Sep. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231218304284

[86] K. Parmar, N. Limbasiya, and M. Dhamecha, "Feature based composite approach for sarcasm detection using MapReduce," in *Proc. 2nd Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Feb. 2018, pp. 587–591.

[87] S. Kannangara, "Mining Twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, New York, NY, USA, Feb. 2018, pp. 751–752, doi: 10.1145/3159652.3170461.

[88] M. J. C. Samonte, C. J. T. Dollete, P. M. M. Capanas, M. L. C. Flores, and C. B. Soriano, "Sentence-level sarcasm detection in English and Filipino tweets," in *Proc. 4th Int. Conf. Ind. Bus. Eng.*, New York, NY, USA, Oct. 2018, pp. 181–186, doi: 10.1145/3288155.3288172.

[89] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of Hindi-English code-mixed social media text for hate speech detection," in *Proc. 2nd Workshop Comput. Model. People's Opinions, Personality, Emotions Social Media*, New Orleans, LA, USA, 2018, pp. 36–41. [Online]. Available: https://www.aclweb.org/anthology/W18-1105

[90] L. Liu, J. L. Priestley, Y. Zhou, H. E. Ray, and M. Han, "A2Text-net: A novel deep neural network for sarcasm detection," in *Proc. IEEE 1st Int. Conf. Cognit. Mach. Intell. (CogMI)*, Dec. 2019, pp. 118–126.

[91] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2506–2515. [Online]. Available: https://www.aclweb.org/anthology/P19-1239

[92] L. H. Son, A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset, "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network," *IEEE Access*, vol. 7, pp. 23319–23328, 2019.

[93] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *Proc. World Wide Web Conf. (WWW)*, New York, NY, USA, 2019, pp. 2115–2124, doi: 10.1145/3308558.3313735.

[94] S. K. Bharti, K. S. Babu, and R. Raman, "Context-based sarcasm detection in Hindi tweets," in *Proc. 9th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Dec. 2017, pp. 1–6.

[95] D. Jain, A. Kumar, and G. Garg, "Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN," *Appl. Soft Comput.*, vol. 91, Jun. 2020, Art. no. 106198. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494620301381

[96] L. Ren, B. Xu, H. Lin, X. Liu, and L. Yang, "Sarcasm detection with sentiment semantics enhanced multi-level memory network," *Neurocomputing*, vol. 401, pp. 320–326, Aug. 2020.

[97] K. Buschmeier, P. Cimiano, and R. Klinger, "An impact analysis of features in a classification approach to irony detection in product reviews," in *Proc. 5th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Baltimore, MD, USA, 2014, pp. 42–49.

[98] E. Filatova, "Sarcasm detection using sentiment flow shifts," in *Proc. Int. Florida Artif. Intell. Res. Soc. Conf. (FLAIRS)*, 2017, pp. 264–269.

[99] M. Khodak, N. Saunshi, and K. Vodrahalli, "A large self-annotated corpus for sarcasm," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018, pp. 641–646. [Online]. Available: https://www.aclweb.org/anthology/L18-1102

[100] R. Justo, T. Corcoran, S. M. Lukin, M. Walker, and M. I. Torres, "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web," *Knowl.-Based Syst.*, vol. 69, pp. 124–133, Oct. 2014.

[101] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "CASCADE: Contextual sarcasm detection in online discussion forums," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 1837–1848. [Online]. Available: https://aclanthology.org/C18-1156

[102] A. D. Dave and N. P. Desai, "A comprehensive study of classification techniques for sarcasm detection on textual data," in *Proc. Int. Conf. Electr., Electron., Optim. Techn. (ICEEOT)*, Mar. 2016, pp. 1985–1991.

[103] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman, "Are word embedding-based features useful for sarcasm detection?" in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 1006–1011. [Online]. Available: https://www.aclweb.org/anthology/D16-1104

[104] M. S. M. Suhaimin, M. H. A. Hijazi, R. Alfred, and F. Coenen, "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*, May 2017, pp. 703–709.

[105] D. Das and A. J. Clark, "Sarcasm detection on Facebook: A supervised learning approach," in *Proc. Int. Conf. Multimodal Interact. Adjunct (ICMI)*, New York, NY, USA, 2018, doi: 10.1145/3281151.3281154.

[106] D. Das, "A multimodal approach to sarcasm detection on social media," M.S. thesis, Missouri State Univ., Springfield, MO, USA, 2019.

[107] J. Eisterhold, S. Attardo, and D. Boxer, "Reactions to irony in discourse: Evidence for the least disruption principle," *J. Pragmatics*, vol. 38, no. 8, pp. 1239–1256, 2006.

[108] R. González-Ibáñez, S. Muresan, and N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Portland, OR, USA, Jun. 2011, pp. 581–586.

[109] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Sep. 2013, pp. 195–198.

[110] E. Riloff, A. Qadir, P. Surve, L. D. Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Seattle, WA, USA, Oct. 2013, pp. 704–714.

[111] T. Ptácek, I. Habernal, and J. Hong, "Sarcasm detection on Czech and English Twitter," in *Proc. 25th Int. Conf. Comput. Linguistics (COLING)*, Aug. 2014, pp. 213–223.

[112] P. Liu, W. Chen, G. Ou, T. Wang, D. Yang, and K. Lei, "Sarcasm detection in social media based on imbalanced classification," in *Web-Age Information Management*, F. Li, G. Li, S.-w. Hwang, B. Yao, and Z. Zhang, Eds. Cham, Switzerland: Springer, 2014, pp. 459–471.

[113] D. K. Tayal, S. Yadav, K. Gupta, B. Rajput, and K. Kumari, "Polarity detection of sarcastic political tweets," in *Proc. Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2014, pp. 625–628.

[114] F. Barbieri, H. Saggion, and F. Ronzano, "Modelling sarcasm in Twitter, a novel approach," in *Proc. 5th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Baltimore, MD, USA, Jun. 2014, pp. 50–58.

[115] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on Twitter," in *Proc. 9th Int. AAAI Conf. Web Social Media*, Jul. 2015, pp. 574–577.

[116] E. Fersini, F. A. Pozzi, and E. Messina, "Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2015, pp. 1–8.

[117] D. Ghosh, W. Guo, and S. Muresan, "Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words," in *Proc. EMNLP*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 1003–1012.

[118] A. Ghosh and D. T. Veale, "Fracking sarcasm using neural network," in *Proc. 7th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, San Diego, CA, USA, 2016, pp. 161–169. [Online]. Available: https://www.aclweb.org/anthology/W16-0425

[119] M. Bouazizi and T. Ohtsuki, "Sarcasm detection in Twitter: 'All your products are incredibly amazing!!!'—Are they really?" in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[120] M. Zhang, Y. Zhang, and G. Fu, "Tweet sarcasm detection using deep neural network," in *Proc. The 26th Int. Conf. Comput. Linguistics: Tech. Papers*, 2016, pp. 2449–2460.

[121] B. Charalampakis, D. Spathis, E. Kouslis, and K. Kermanidis, "A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets," *Eng. Appl. Artif. Intell.*, vol. 51, pp. 50–57, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197616000117

[122] D. I. H. Farías, V. Patti, and P. Rosso, "Irony detection in Twitter: The role of affective content," *ACM Trans. Internet Technol.*, vol. 16, no. 3, pp. 1–24, Aug. 2016, doi: 10.1145/2930663.

[123] A. Joshi, P. Bhattacharyya, M. Carman, J. Saraswati, and R. Shukla, "How do cultural differences impact the quality of sarcasm annotation?: A case study of Indian annotators and American text," in *Proc. 10th SIGHUM Workshop Lang. Technol. Cultural Heritage, Social Sci., Humanities*, 2016, pp. 95–99.

[124] A. Joshi, P. Jain, P. Bhattacharyya, and M. Carman, "Who would have thought of that!': A hierarchical topic model for extraction of sarcasm-prevalent topics and sarcasm detection," in *Proc. Workshop Extra-Propositional Aspects Meaning Comput. Linguistics (ExProM)*. Osaka, Japan, Dec. 2016, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/W16-5001

[125] D. Al-Ghadhban, E. Alnkhilan, L. Tatwany, and M. Alrazgan, "Arabic sarcasm detection in Twitter," in *Proc. Int. Conf. Eng. MIS (ICEMIS)*, May 2017, pp. 1–7.
[126] P. Dharwal, T. Choudhury, R. Mittal, and P. Kumar, "Automatic sarcasm detection using feature selection," in *Proc. 3rd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Dec. 2017, pp. 29–34.
[127] R. K. Gupta and Y. Yang, "CrystalNest at SemEval-2017 task 4: Using sarcasm detection for enhancing sentiment classification and quantification," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, Vancouver, BC, Canada, 2017, pp. 626–633. [Online]. Available: https://www.aclweb.org/anthology/S17-2103
[128] A. Mishra, D. Kanojia, S. Nagar, K. Dey, and P. Bhattacharyya, "Harnessing cognitive features for sarcasm detection," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, vol. 1, 2016, pp. 1095–1104. [Online]. Available: https://www.aclweb.org/anthology/P16-1104
[129] A. Ghosh and T. Veale, "Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 482–491. [Online]. Available: https://www.aclweb.org/anthology/D17-1050
[130] S. Saha, J. Yadav, and P. Ranjan, "Proposed approach for sarcasm detection in Twitter," *Indian J. Sci. Technol.*, vol. 10, no. 25, pp. 1–8, 2017.
[131] P. Deshmukh and S. Solanke, "Review paper: Sarcasm detection and observing user behavioral," *Int. J. Comput. Appl.*, vol. 166, no. 9, pp. 39–41, May 2017. [Online]. Available: http://www.ijcaonline.org/archives/volume166/number9/27701-2017914119
[132] S. Mukherjee and P. K. Bala, "Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering," *Technol. Soc.*, vol. 48, pp. 19–27, Feb. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0160791X16300070
[133] T. Jain, N. Agrawal, G. Goyal, and N. Aggrawal, "Sarcasm detection of tweets: A comparative study," in *Proc. 10th Int. Conf. Contemp. Comput. (IC3)*, Aug. 2017, pp. 1–6.
[134] A. Joshi, D. Kanojia, P. Bhattacharyya, and M. J. Carman, "Sarcasm suite: A browser-based engine for sarcasm detection and generation," in *Proc. AAAI*, 2017, pp. 5095–5096.
[135] A. G. Prasad, S. Sanjana, S. M. Bhat, and B. S. Harish, "Sentiment analysis for sarcasm detection on streaming short text data," in *Proc. 2nd Int. Conf. Knowl. Eng. Appl. (ICKEA)*, Oct. 2017, pp. 1–5.
[136] D. Ghosh, A. Richard Fabbri, and S. Muresan, "The role of conversation context for sarcasm detection in online interactions," in *Proc. 18th Annu. SIGdial Meeting Discourse Dialogue*, Saarbrücken, Germany, 2017, pp. 186–196. [Online]. Available: https://www.aclweb.org/anthology/W17-5523
[137] L. Peled and R. Reichart, "Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, vol. 1, 2017, pp. 1690–1700. [Online]. Available: https://www.aclweb.org/anthology/P17-1155
[138] H. Elgabry, S. Attia, A. Abdel-Rahman, A. Abdel-Ate, and S. Girgis, "A contextual word embedding for Arabic sarcasm detection with random forests," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, Apr. 2021, pp. 340–344. [Online]. Available: https://aclanthology.org/2021.wanlp-1.43
[139] D. Faraj, D. Faraj, and M. Abdullah, "Sarcasmdet at sarcasm detection task 2021 in Arabic using AraBERT pretrained model," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, Apr. 2021, pp. 345–350. [Online]. Available: https://aclanthology.org/2021.wanlp-1.44
[140] H. Nayel, E. Amer, A. Allam, and H. Abdallah, "Machine learning-based model for sentiment and sarcasm detection," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, Apr. 2021, pp. 386–389. [Online]. Available: https://aclanthology.org/2021.wanlp-1.51
[141] M. Shrivastava and S. Kumar, "A pragmatic and intelligent model for sarcasm detection in social media text," *Technol. Soc.*, vol. 64, Feb. 2021, Art. no. 101489. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X20312926
[142] A. El Mahdaouy, A. El Mekki, K. Essefar, N. El Mamoun, I. Berrada, and A. Khoumsi, "Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language," 2021, *arXiv:2106.12488*.
[143] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, *Affective Dependency Graph for Sarcasm Detection*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1844–1849. [Online]. Available: https://doi.org/10.1145/3404835.3463061
[144] F. Husain and O. Uzuner, "Leveraging offensive language for sarcasm and sentiment detection in Arabic," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, Apr. 2021, pp. 364–369. [Online]. Available: https://aclanthology.org/2021.wanlp-1.47
[145] M. S. Razali, A. A. Halin, L. Ye, S. Doraisamy, and N. M. Norowi, "Sarcasm detection using deep learning with contextual features," *IEEE Access*, vol. 9, pp. 68609–68618, 2021.
[146] F. Yao, X. Sun, H. Yu, W. Zhang, W. Liang, and K. Fu, "Mimicking the brain's cognition of sarcasm from multidisciplines for Twitter sarcasm detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 13, 2021, doi: 10.1109/TNNLS.2021.3093416.
[147] X. Guo, B. Li, H. Yu, and C. Miao, "Latent-optimized adversarial neural transfer for sarcasm detection," 2021, *arXiv:2104.09261*.
[148] A. Kamal and M. Abulaish, "CAT-BiGRU: Convolution and attention with bi-directional gated recurrent unit for self-deprecating sarcasm detection," *Cognit. Comput.*, vol. 14, no. 1, pp. 91–109, Jan. 2022.
[149] Y. Du, T. Li, M. S. Pathan, H. K. Teklehaimanot, and Z. Yang, "An effective sarcasm detection approach based on sentimental context and individual expression habits," *Cognit. Comput.*, vol. 14, no. 1, pp. 78–90, Jan. 2022.
[150] R. Ghaeini, X. Z. Fern, and P. Tadepalli, "Attentional multi-reading sarcasm detection," 2018, *arXiv:1809.03051*.
[151] Y. Zhang, Y. Liu, Q. Li, P. Tiwari, B. Wang, Y. Li, H. M. Pandey, P. Zhang, and D. Song, "CFN: A complex-valued fuzzy network for sarcasm detection in conversations," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3696–3710, Dec. 2021.
[152] R. Rakov and A. Rosenberg, "'Sure, I did the right thing': A system for sarcasm detection in speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Jan. 2013, pp. 842–846.
[153] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, "Multi-modal sarcasm detection and humor classification in code-mixed conversations," *IEEE Trans. Affect. Comput.*, early access, May 26, 2021, doi: 10.1109/TAFFC.2021.3083522.
[154] A. Kumar, S. Dikshit, and V. H. C. Albuquerque, "Explainable artificial intelligence for sarcasm detection in dialogues," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–13, Jul. 2021.
[155] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE MultimediaMag.*, vol. 28, no. 2, pp. 86–95, Apr. 2021.
[156] Z. Wen, L. Gui, Q. Wang, M. Guo, X. Yu, J. Du, and R. Xu, "Sememe knowledge and auxiliary information enhanced approach for sarcasm detection," *Inf. Process. Manage.*, vol. 59, no. 3, May 2022, Art. no. 102883. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457322000139
[157] S. M. Mohammad, S. Kiritchenko, and J. Martin, "Identifying purpose behind electoral tweets," in *Proc. 2nd Int. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, New York, NY, USA, 2013, pp. 1–9.
[158] S. Oraby, V. Harrison, L. Reed, E. Hernandez, E. Riloff, and M. Walker, "Creating and characterizing a diverse corpus of sarcasm in dialogue," in *Proc. 17th Annu. Meeting Special Interest Group Discourse Dialogue*, Los Angeles, CA, USA, 2016, pp. 31–41. [Online]. Available: https://www.aclweb.org/anthology/W16-3604
[159] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King, "A corpus for research on deliberation and debate," in *Proc. LREC*, Istanbul, Turkey, vol. 12, 2012, pp. 812–817.
[160] I. Abu Farha and W. Magdy, "Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, Apr. 2021, pp. 21–31. [Online]. Available: https://aclanthology.org/2021.wanlp-1.3
[161] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *LREC*, vol. 10, 2010, pp. 2200–2204.
[162] H. Zhang and D. Li, "Naïve Bayes text classifier," in *Proc. IEEE Int. Conf. Granular Comput. (GRC)*, Nov. 2007, p. 708.
[163] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory (COLT)*, 1992, pp. 144–152.
[164] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[165] G. Lakoff and M. Johnson, *Metaphors we Live by*. Chicago, IL, USA: Univ. Chicago Press, 2008.
[166] M. A. Di Gangi, G. Lo Bosco, and G. Pilato, "Effectiveness of data-driven induction of semantic spaces and traditional classifiers for sarcasm detection," *Natural Lang. Eng.*, vol. 25, no. 2, pp. 257–285, Mar. 2019.
[167] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling," in *Proc. LREC*, vol. 6, 2006, pp. 1222–1225.
[168] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Apr. 1995.
[169] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[170] C. Manning, "MaXENT models and discriminative estimation," Natural Lang. Process., Stanford Univ., Stanford, CA, USA, Lect. Notes CS224N, 2005.
[171] J. S. Cramer, "The origins of logistic regression," Tinbergen Inst., Amsterdam, The Netherlands, Tinbergen Institute Discussion Papers 02-119/4, Dec. 2002. [Online]. Available: https://ideas.repec.org/p/tin/wpaper/20020119.html
[172] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[173] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.
[174] H. Daumé, J. Langford, and D. Marcu, "Search-based structured prediction," *Mach. Learn.*, vol. 75, no. 3, pp. 297–325, Jun. 2009.
[175] S. Castro, D. Hazarika, V. Perez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," 2019, *arXiv:1906.01815*.
[176] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[177] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[178] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[179] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-10590-1_53#citeas
[180] M. Auli, M. Galley, C. Quirk, and G. Zweig, "Joint language and translation modeling with recurrent neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Seattle, WA, USA, Oct. 2013, pp. 1044–1054. [Online]. Available: https://aclanthology.org/D13-1106
[181] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep reinforcement learning for dialogue generation," 2016, *arXiv:1606.01541*.
[182] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.
[183] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[184] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.
[185] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[186] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, vol. 1, 2018, pp. 328–339. [Online]. Available: https://aclanthology.org/P18-1031
[187] S. Merity, N. Shirish Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," 2017, *arXiv:1708.02182*.

MONDHER BOUAZIZI (Member, IEEE) received the B.E. degree in communications from SUP-COM, Carthage University, Tunisia, in 2010, and the M.E. and Ph.D. degrees from Keio University, in 2017 and 2019, respectively. He worked as a Telecommunication Engineer (access network quality and optimization) for three years with Ooredoo Tunisia (Ex. Tunisiana). He currently works as a Specially Appointed Assistant Professor with the Ohtsuki Laboratory, Department of Information and Computer Science, Faculty of Science and Technology, Keio Univeristy. He has published several journal and international conference papers. He is engaged in research on machine learning, deep learning, data mining, sensors, and signal processing.

He is a member of ACM and IEICE. He has received the Telecommunications Advancement Foundation Student Award 2016, the IEEE/ACM ICSIM 2021 Best Paper Award, and the A3 Workshop 2021 Best Presentation Award.

TOMOAKI OHTSUKI (OTSUKI) (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1990, 1992, and 1994, respectively.

From 1994 to 1995, he was a Postdoctoral Fellow and a Visiting Researcher in electrical engineering with Keio University. From 1993 to 1995, he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005, he was with the Science University of Tokyo. From 1998 to 1999, he was with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley. In 2005, he joined Keio University, where he is currently a Professor. He has published more than 215 journal articles and 415 international conference papers. His research interests include wireless communications, optical communications, signal processing, and information theory. He is a Distinguished Lecturer of the IEEE, a fellow of the IEICE, and a member of the Engineering Academy of Japan. He was a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, the Ericsson Young Scientist Award 2000, the 2002 Funai Information and Science Award for Young Scientist, the IEEE the 1st Asia–Pacific Young Researcher Award 2001, the 5th International Communication Foundation (ICF) Research Award, the 2011 IEEE SPCE Outstanding Service Award, the 27th TELECOM System Technology Award, the *ETRI Journal*'s 2012 Best Reviewer Award, and 9th International Conference on Communications and Networking in China 2014 (CHINACOM'14) Best Paper Award. He served as the Chair for the IEEE Communications Society and the Signal Processing for Communications and Electronics Technical Committee. He has served as the General-Co Chair, the Symposium Co-Chair, and the TPC Co-Chair for many conferences, including IEEE GLOBECOM 2008, SPC, IEEE ICC 2011, CTS, IEEE GLOBECOM 2012, SPC, IEEE ICC 2020, SPC, IEEE APWCS, IEEE SPAWC, and IEEE VTC. He gave tutorials and keynote speeches at many international conferences, including IEEE VTC, IEEE PIMRC, and IEEE WCNC. He was the Vice President and the President of the Communications Society of the IEICE. He has served as the Technical Editor for the *IEEE Wireless Communications* magazine and an Editor for *Physical Communication* (Elsevier). He is currently serving as an Area Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.