Text-based Sarcasm Detection on Social Networks: A Systematic Review

Amal Alqahtani[1], Lubna Alhenaki[2], Abeer Alsheddi[3]
Computer Science Department, King Khalid University, Abha, Saudi Arabia[1]
Computer Science Department, Majmaah University, Al-Majmaah, Saudi Arabia[2]
Computer Science Department, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia[3]
King Saud University, Riyadh, Saudi Arabia[1, 3]

***Abstract*—Sarcasm is a sophisticated phenomenon used for conveying a meaning that differs from what is being said, and it is usually used to express displeasure or ridicule others. Sentiment analysis is a process of uncovering the subjective information from a text. Detecting figurative language such as irony or sarcasm, is a focused challenging research field of sentiment analysis. Detecting and understanding the use of sarcasm in social networks could provide businesses and politicians with significant insight, since it reflects people's opinions about certain topics, news, and products. This has especially become relevant recently because sarcastic texts have been trending on social networks and are being posted by millions of active users. As a result of this situation, there is now an increasing amount of research on the detection of sarcasm in social network posts. Many works have been published on sarcasm detection, and they include a wide variety of techniques based on rules, lexicons, traditional machine learning, deep learning, and transformers. However, sarcasm detection is a challenging task due to the ambiguity and non-straightforward nature of sarcastic text. In addition, very few reviews have been conducted on the research in this area. Therefore, this systematic review mainly aims at exploring the newly published sarcasm detection articles on social networks in the years between 2019 and 2022. Several databases were extensively searched, and 30 articles that met the criteria were included. The selected articles were reviewed based on their approaches, datasets, and evaluation metrics. The findings emphasized that deep learning is the most commonly used technique for sarcasm detection in recent literature, and Twitter and F-measure are the most used source and performance metric, respectively. Finally, this article presents a brief discussion regarding the challenges in sarcasm detection and future research directions.**

***Keywords*—Sentiment analysis; figurative language; sarcasm detection; irony; machine learning; deep learning; transformer**

I. Introduction

Over the last few years, natural language processing (NLP) has been one of the most active areas of artificial intelligence (AI) research. Researchers in this area have made considerable effort to enable machines to mimic the human ability for language, and the results have often been ground-breaking. For example, sentiment analysis, also known as opinion mining, is an NLP task that involves identifying the subjectivity and sentiments present in opinions [1]. Social networks, such as Twitter and Facebook, are gaining increasing popularity and have millions of active users. In particular, Twitter is one of the most popular social networks that attracts millions of users [2]. In addition, text is considered as the most commonly used form of communication, with social network posts varying from short-text data, such as tweets, to long-text posts such as debates.

Sarcasm can be defined as saying or writing the opposite of what is intended. As a result, sarcasm generates ambiguous and non-straightforward data. For instance, "I love to go to the dentist!" is an obvious example of the use of sarcasm for expressing negative feelings. Overall, it is occasionally hard to efficiently recognize sarcasm due to the contradiction between the implicit and explicit meaning [3]. Moreover, textual sarcasm is challenging due to the lack of tone and facial expressions, and this makes it hard for even human beings to detect sarcasm [4]. Therefore, textual sarcasm is a vague task that needs to be studied carefully. A well-designed NLP model for text-based sarcasm detection is, thus, crucial.

Over the past years, a few reviews about sarcasm detection in social networks have been published, but most of them focused mainly on the implementation phase, for example, [5],[6] and [7]. However, some of the previous research did not cover all the approaches used for sarcasm detection. For example, the authors in [5] reviewed and analyzed machine learning-based sarcasm detection studies and found that support vector machine (SVM) is the most frequently utilized classification algorithm for sarcasm detection. However, there are many other techniques in use that need to be studied. The researchers in [7] reviewed the rule-based, statistical-based, and deep learning (DL) approaches for sarcasm detection but did not consider other popular approaches such as transformers, while the researchers in [6] only presented a technical review of sarcasm detection algorithms and reported the mostly frequently used algorithms for sarcasm identification.

Based on the gaps in the literature discussed above, the main aim of this article is to conduct a systematic literature review (SLR) that focuses on identifying and analyzing text-based sarcasm detection articles on social networks based on their development approaches, evaluation metrics, and datasets. Moreover, this article presents an overview of the main sarcasm detection challenges and future possible improvements. To achieve these objectives, the following four research questions will be answered:

- RQ1: What are the main approaches used for the development of automatic sarcasm detection models?

- RQ2: What are the most commonly used metrics to evaluate the performance of sarcasm detection models?
- RQ3: What datasets are most commonly used for detecting sarcasm on social networks?
- RQ4: What are the main challenges in sarcasm detection?

The remainder of this article is organized as follows. Section II provides the problem statement, and Section III describes the methodology used in this SLR. The approaches, metrics, and datasets of the reviewed articles are provided in Sections IV, V, and VI, respectively. Section VII discusses the findings, research problems, and future research directions. Finally, the conclusion is provided in Section VIII.

II. Problem Statement

Over the past decade, the increase in the number of social network users has caused researchers to deeply investigate and analyze data on social networks. Sarcasm detection is one of the most challenging tasks and is a hot topic in the NLP field. Non-straightforward sarcastic data may reflect positive or negative sentiments or both polarities. In fact, it is difficult to detect sarcasm because sarcastic text is often obscure and ambiguous. In other words, there is little agreement on the actual intention behind indirect sarcastic sentences even by humans, and this makes it even harder to accomplish such tasks with AI technology. Most of the text-based sarcasm cannot be interpreted literally since the actual purpose of the sarcastic text might be the opposite of the apparent meaning of the text. Moreover, the lack of body language and voice tone in text-based sarcasm make it difficult to understand sarcasm in text. Another challenge to sarcasm detection is that the context of sarcasm is strongly dependent on cultures, personalities, and languages.

Sarcasm detection is important for tracking people's opinion and satisfaction in relation to products. Therefore, sarcasm detection is an essential task for decision making by businesses. Social networks, by nature, are rich in sarcastic texts, and this further increases the need for extensive analysis and study. However, applying basic sentiment techniques such as rule-based techniques, with sarcastic text is not sufficient. Therefore, there is a strong need for a well-designed model specifically oriented towards sarcasm detection tasks. The availability of recent review in sarcasm detection field would pave the way for a new novel solution. Therefore, it is crucial to conduct a review that covers the most recent techniques as well as the state-of the art techniques.

Recently, several works have been published on sarcasm detection with machine-learning (ML), DL, and transformer techniques. However, a limited number of the reviews so far have conducted in-depth investigations into sarcasm detection. Therefore, the present SLR comprehensively covers recent articles on text-based sarcasm detection in social networks that were published between 2019 and 2022. In addition, the reviews published so far, that is [8], [9], [10], [11] and [7] have several limitations. For instance, the study in [8] used a different database and selection criteria compared to this study, and the studies in [9] and [10] differ with regard to their research questions. Further, the challenges involved in the development of an effective model for sarcasm detection are not highlighted in [11]. The researchers in [7] did not provide sufficiently detailed characteristics and findings regarding the recent sarcasm datasets and metrics. To sum up, this SLR was conducted with the aim of filling in the highlighted gaps in the previous reviews, as described above. With this survey, our aim is to identify and analyze text-based sarcasm detection articles on social networks based on their development approaches, evaluation metrices, and datasets.

III. Survey Methodology

This SLR uses the Kitchenham guidelines for reviewing articles on sarcasm detection [12]. According to these guidelines, the three stages of a review are planning, conducting, and reporting the review. The following subsections provide the details of these three stages. First, Section A presents the planning stage, including the goals and research questions, database identification and search procedure, and inclusion and exclusion criteria. Second, article selection and quality assessment. Third, from Section IV to Section VI the third stage is reported.

A. Planning

1) Goals and research questions: The primary purpose of this SLR is to identify and analyze articles on the state of the art of sarcasm detection tools based on their development approaches, evaluation metrics, most commonly used datasets, and the major challenges to sarcasm detection identified. To achieve these objectives, the following research questions are investigated:

- RQ1: What are the main approaches used for automatic sarcasm detection models?
- RQ2: What are the commonly used metrics to evaluate the performance of sarcasm detection models?
- RQ3: What datasets are most commonly used for detecting sarcasm on social networks?
- RQ4: What are the main challenges in sarcasm detection?

2) Databases identification and search procedure: Four scientific databases, namely, IEEE, Springer, ScienceDirect, and ACM, were used to search and identify relevant research articles. The search was conducted using nine keywords based on specific selection criteria, which will be described in Section 3. The keywords were selected based on those mentioned in [13],[8] and [9]. Table I presents the number of selected articles based on the keywords and database names.

3) Inclusion and exclusion criteria: The inclusion and exclusion criteria for selecting the most relevant articles were based on the objectives of this SLR. The inclusion criteria were as follows:

a) Articles published in the English language 2. Articles published from 2019 to 2022.

b) Full journal articles.

c) Articles published in the field of computer science.

TABLE I. NUMBER OF SELECTED ARTICLES BASED ON THE KEYWORDS FOR EACH OF THE FOUR DATABASES

Keyword	IEEE	Springer	ScienceDirect	ACM
Sarcasm AND Detection AND Sentiment analysis	11	9	150	83
Sarcasm AND Detection AND Artificial intelligence	12	8	66	48
Sarcasm AND Detection AND machine learning	7	9	155	82
Sarcasm AND Detection AND Deep learning	9	9	139	82
Sarcasm AND Recognition AND Sentiment analysis	2	9	87	83
Sarcasm AND Recognition AND Artificial intelligence	4	8	44	47
Sarcasm AND Recognition AND machine learning	1	9	85	81
Sarcasm AND Recognition AND Deep learning	1	9	84	81
Irony AND Detection AND Sentiment analysis	1	1	92	56
Irony AND Detection AND Artificial intelligence	0	1	47	37
Irony AND Detection AND machine learning	0	1	100	54
Irony AND Detection AND Deep learning	0	1	92	56
Irony AND Recognition AND Sentiment analysis	0	2	62	0
Irony AND Recognition AND Artificial intelligence	2	1	37	41
Irony AND Recognition AND machine learning	2	2	54	0
Irony AND Recognition AND Deep learning	2	0	47	0
Figurative language AND Detection AND Sentiment analysis	4	4	37	12
Figurative language AND Detection AND Artificial intelligence	3	0	16	11
Figurative language AND Detection AND machine learning	1	1	45	12
Figurative language AND Detection AND Deep learning	3	3	44	12
Figurative language AND Recognition AND Sentiment analysis	0	4	17	13
Figurative language AND Recognition AND Artificial intelligence	1	0	14	52
Figurative language AND Recognition AND machine learning	0	2	30	17
Figurative language AND Recognition AND Deep learning	1	3	30	17
Total	67	96	1574	977

A large number of articles met the inclusion criteria, and these were filtered using the following three exclusion criteria.

- Titles and abstracts that were irrelevant to sarcasm detection.
- Duplication.
- Inability of the articles to address the research questions.

As the number of articles retrieved was too large to process manually, it is assumed that the retrieved articles in a database search engine are arranged in accordance with the keywords. According to the first exclusion criterion, articles with titles and abstracts that were not related to sarcasm detection were excluded. Next, duplicate articles that appear in more than one of the databases were excluded. The last criterion relates to whether the articles could address the research questions and involves quality assessment of the candidate articles, as discussed in the following subsection.

4) Article selection: The initial search in the databases returned about 2726 articles. Table I details the number of articles returned for each possible keyword query in all four databases. In general, the maximum number of articles (1574) was retrieved from ScienceDirect database; this is probably due to differences in the content of the databases, interests, and domains. Moreover, the highest number of articles was retrieved with the query "Sarcasm AND Detection AND Machine learning".

For screening the retrieved articles, the inclusion and exclusion criteria described in the previous subsection are applied. Based on these criteria, 2634 irrelevant articles were excluded, and 92 relevant articles were considered. Following this, 47 duplicated articles were further excluded, and the remaining 45 articles were considered for deeper investigation. Finally, 15 articles that did not address the research questions were excluded, and this left us with 30 articles. Fig. 1 illustrates the article selection process.

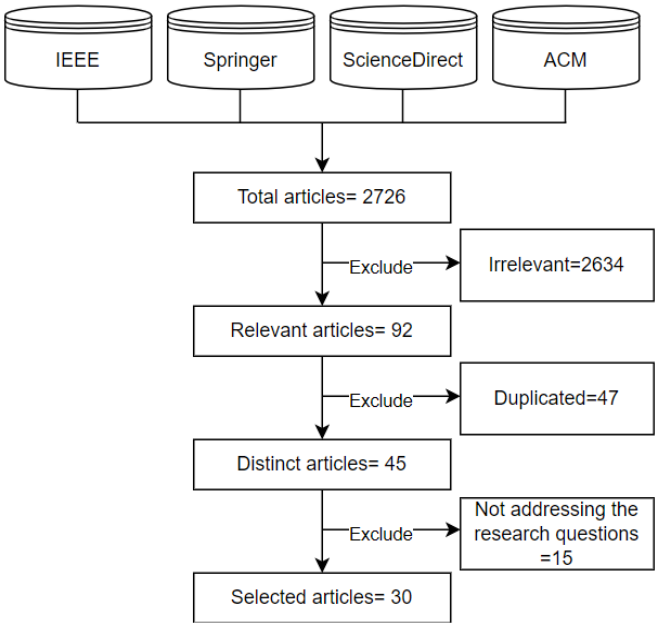

Fig. 1. Flowchart depicting the article selection process.

5) Quality assessment: This section describes quality assessment of the articles based on the method described in [14]. The articles were assessed using the following 10 questions, and articles for which the response was "yes" for at least seven questions were selected.

- Are the article objectives clearly defined?
- Does the article provide a brief description of the previous sarcasm detection approaches?
- Are the evaluation metrices explained clearly?
- Is the article structure designed appropriately?
- Are the data collection processes explained in detail?
- Are the approach, formulation, and analysis described adequately?
- Does the article list the used dataset?
- Is the article understandable and well-written?
- Does the article utilize a well-designed methodology?
- Does the article present and interpret the results clearly?

IV. Sarcasm Detection Approaches and Techniques

There are many studies on NLP methods for sarcasm detection. Recent articles in the field of text-based sarcasm detection on different social networking platforms and online media is surveyed and discussed in this section, but it is not meant to be exhaustive. Sarcasm detection approaches can be categorized based on the classification technique into rule-based, lexicon-based, traditional ML-based, DL-based, and transformer-based approaches. Fig. 2 presents the general structure of sarcasm detection approaches along with their common techniques in the selected articles.

The related works are categorized into five subsections based on the approaches they have explored: Section A focuses on the rule-based approach; Section B, the lexicon-based approach; Section C, traditional ML-based approaches; Section D, DL-based approaches; Section E, the transformer-based approaches. Table II presents a detailed comparison of these works. Overall, traditional ML, DL, and transformer-based approaches are becoming popular in the field of NLP, especially in the area of sarcasm detection. Therefore, in this SLR, studies that focus on these three approaches will be studied in detail.

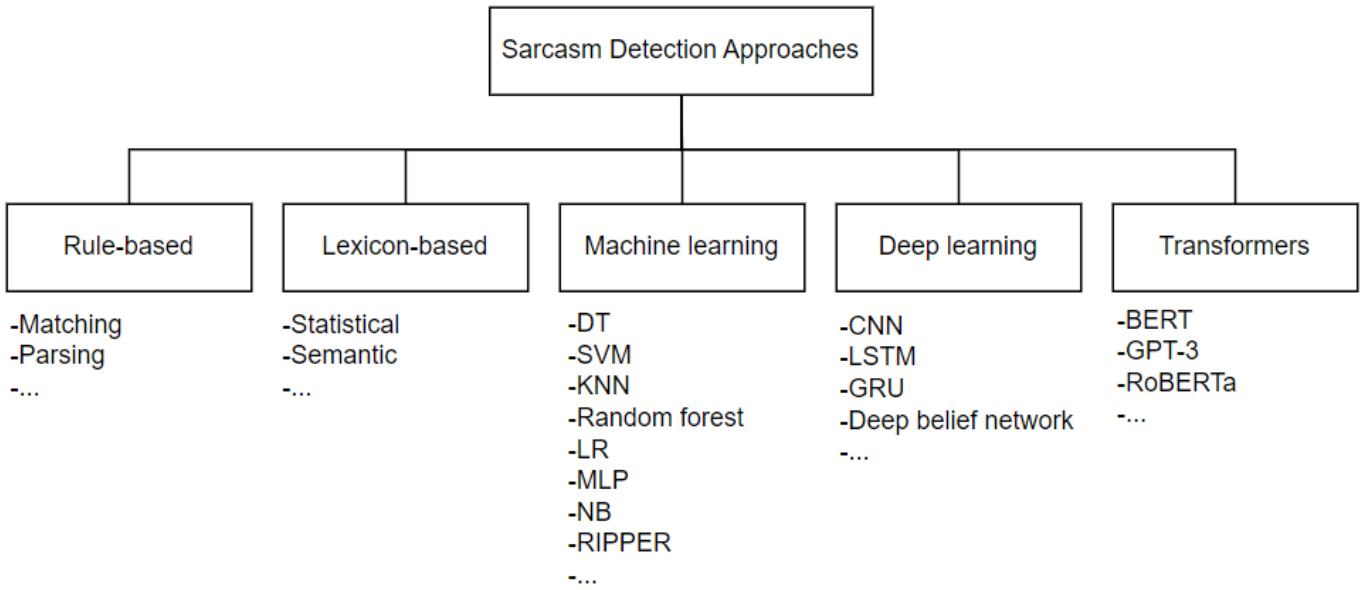

Fig. 2. General structure of sarcasm detection approaches.

TABLE II. SUMMARY OF THE REVIEWED ARTICLES

No.	Article	Model	Year
1	[33]	Combination of a machine learning classifier	2021
2	[31]	Combination of a machine learning classifier	2020
3	[34]	Combination of a machine learning classifier	2021
4	[36]	Combination of a machine learning classifier	2020
5	[32]	Combination of a machine learning classifier	2022
6	[38]	SVM classifier	2022
7	[40]	Bi-LSTM	2019
8	[41]	Att-BiLSTM and convNet deep learning model	2019
9	[42]	MHA-BiLSTM (Multi-Head Attention-based Bidirectional Long Short-term Memory)	2020
10	[43]	CNN	2020
11	[44]	MMNSS (Multi-level Memory Network based on sentiment semantics)	2020
12	[45]	Deep learning approach that consists of an input, embedding, convolutional, Bi-directional Gated Recurrent Unit (BiGRU), and two attention layers	2022
13	[46]	Term-weighted word embedding combined with trigram and 3-layer LSTM	2021
14	[47]	CNN + attention-based BiLSTM	2022
15	[48]	Pre-trained (BERT) for word embeddings + GRU	2021
16	[49]	Multi-task Bi-GRU and attention-based CNN	2021
17	[50]	Deep belief network	2022
18	[51]	HA-LSTM (hybrid attention-based Long Short-Term Memory)	2021
19	[52]	Combination of a machine learning classifier and CNN	2021
20	[53]	BERT, CNN, and LSTM	2022
21	[10]	AWD-LSTM (Averaged Stochastic Gradient Descent Weight-Dropped LSTM)	2022
22	[54]	Attention-based BiGRU	2022
23	[55]	Combination of a machine learning classifier and LSTM	2021
24	[59]	BERT, BiLSTM, and NetXtVLAD	2020
25	[60]	BERT	2021
26	[63]	Bi-LSTM, BERT, and GloVe	2021
27	[66]	BERT	2022
28	[67]	COMET Model	2021
29	[70]	RCNN-RoBERTa	2019
30	[71]	Encoder model called LMTweets with multiple techniques	2021

A. Rule-based Approach

This approach comprises a set of predefined human-made rules that act as indicators of sarcasm. Different researchers have proposed different approaches for making the rules such as parsing and matching. For example, some authors used hashtags as a key indicator of sarcasm. That is, they assumed that if tweets contain specific hashtags and do not fit in with the rest of the tweets, then that statement is sarcasm [15]. Another author combined two rule-based approaches: the first one is used for developing and recognizing the parse tree, and the other approach captures hyperboles features by using interjection and intensifiers together [16]. A third rule-based approach is "simile," which involves comparing two things directly. One of the studies that utilized this approach for sarcasm detection was described in [17].

B. Lexicon-based Approach

Lexicon-based approaches rely on a predefined collection of words, referred to as a lexicon, with each of the words assigned to a particular polarity category indicating its nature, namely, positive, natural, negative, which are represented by the numerical values -1, 0, and +1, respectively. The lexicon can be weighted or unweighted, such that the words which induce higher positivity or negativity are given a higher probability [1]. In this sarcasm detection process, a bags-of-lexicon which comprises a positive sentiment, a negative sentiment, a positive context, and a negative context is created. A text is divided into tokens of a single word, and the score of each token is obtained using the lexicon. The overall score of the text is determined by adding the individual scores and calculating the average, which is used to determine the sentiment of the text [18]. Sarcasm is detected when a positive context comprises a negative sentiment or a negative context comprise a positive sentiment [16]. An advantage of the lexicon-based approach is that it is suitable at both the sentence and feature level. Moreover, it can be considered as an unsupervised approach because it does not include a training process. However, a major limitation is that it is domain dependent, as the same word would have different meanings according to its context. For example, the word "small" in the statement "this camera is extremely small" could imply a positive sentiment, whereas the use of "small" in "the TV screen is too small" implies a negative sentiment. This could be overcome by constructing a domain-specific lexicon or adapting the current lexicons [19]. In addition, lexicon-based approaches can be divided into the corpus-based approach and the dictionary-based approach.

1) Corpus-based approaches: The corpus-based approach starts with a pre-defined list of polar words with their orientation; their syntactic and co-occurrence pattern is then investigated to obtain other polar words and their corresponding orientation to obtain a bigger "corpus". This approach was first proposed in [20]: a list of adjectives (polar words) with their orientation were pre-defined, and new adjectives and orientations were added using linguistic constraints and rules. For example, in the sentence "the question is simple and easy," there is a connective word "AND" which indicates that both adjectives have the same orientation; in contrast, the connective word "OR" indicates that the adjectives have opposite orientations. This approach is known as "sentiment consistency".

There are two approaches to determining the orientation of polar words, namely, the statistical approach and the semantic approach [18]. The statistical approach relies on the notion that words with similar orientation are likely to appear together frequently. Hence, the new unknown word can be assigned a certain orientation based on its frequency and co-occurrence with other words for which the orientation is known [21]. Some studies on the statistical approach have been published, such as [22] and [23]. The semantic approach, on the other hand, exploits the sentiment dictionary to discover synonyms and antonyms in order to construct a lexicon that can be used to assign the same orientation to words that are semantically similar [24]. Some studies have utilized the semantic approach to build the lexicon, such as [25] and [26]. In addition, a hybrid method can be used to take advantage of both approaches, as described in the work of Zhang [27].

2) Dictionary-based approaches: The dictionary-based approach is roughly based on the idea that synonymous words have the same orientation, and antonyms have the opposite orientation. Therefore, an initial well-known dictionary, such as Thesauri, is constructed with a pre-defined lists of polar words and their orientation. Then, this dictionary is expanded manually based on synonyms and antonyms of the existing words by adding new words and their orientation iteratively until no more words can be added [28]. Finally, manual evaluation and correction can be performed to ensure the validity of the dictionary. This is known as the bootstrapping technique. A popular recently developed dictionary is SentiWordNet 3.0, which uses the automatic annotation of Synsets of WordNet 3 [29]. In addition, Park and Kim in [30] proposed a rule-based method to label the words in advertisements based on three online dictionaries.

C. Traditional ML-based Approaches

Since the earlier years, many studies on text sarcasm detection utilized supervised ML classifiers. Based on the surveyed studies, SVM is one of the most popular classifiers, as evident in [31], [32] and [33].

In 2020, researchers in [31] proposed a sarcasm type detection approach that utilized the multi-rule based ensemble feature selection model. The main aim of this study was to determine the level of hurt that is expressed in sarcasm. Four classes of sarcasm type were determined, including rude, raging, polite, and deadpan. This study used ensemble learning to identify the optimal feature set among all the features and to classify a tweet as sarcastic or not. Following this, the type of sarcasm was determined by using a rule-based approach. This experiment was conducted by using tweets obtained through the Twitter Application Programming Interface (API) Tweepy and Twython. A study conducted in 2021 [33] developed three kinds of ensemble classification algorithms for detecting sarcasm with the Principal Component Analysis (PCA) algorithm. The ensemble classification algorithm is a combination of SVM, KNN, decision tree, logistic regression, and Multi-layer Perceptron (MLP). The three models were tested on five datasets of different sizes from the Twitter streaming API.

Another related study [34] used different ML techniques, such as SVM and logistic regression, for classification. The main contribution was combining the features extracted from a Convolutional Neural Network (CNN) architecture with contextual handcrafted features to obtain the most optimal features. The experiments were conducted on a Twitter dataset created by the researchers and shared publicly [35]. One of the studies that utilized the supervised ML classifier approach with BERT and GloVe embeddings for sarcasm identification [36] also used a Twitter dataset for evaluation. A related study [32] investigated tweets with a negative mood and hyperboles to detect sarcasm. Several ML algorithms, such as SVM, random forest (RF), and RF with bagging, were utilized to analyze five hyperbole features, namely, interjection, intensifier, capital letter, punctuation mark, and elongated word. This study was conducted on tweets collected using the Twitter streaming API [37].

In 2022, the researchers in [38] proposed an intelligent ML-based sarcasm detection and classification (IMLB-SDC) technique in which an SVM classifier is used for sarcasm identification on social networks. The proposed model consists of different stages, namely, preprocessing, feature engineering, feature selection and classification, and parameter tuning.

D. DL-based Approaches

DL is gaining more attention in the sarcasm detection process, since it can be used to obtain better results from unstructured data. It has the ability to learn from a given text in order to either extract automated features or perform sarcasm classification. Based on our investigations, most sarcasm detection articles combine several DL techniques in a model. The most frequently used DL approaches are CNN, artificial neural network, and long short-term memory (LSTM). These are described below.

CNN is a version of the feed forward neural network with multiple hidden layers. It first emerged in computer vision applications, and since then, it has been widely used recently in NLP applications. The network comprises an input layer, hidden layers that consist of many convolution layers, pooling layers, normalization layers, a fully connected layer, and an output layer. The generic workflow of CNN in sarcasm detection is as follows: The convolution layer extracts the features from the input text (word embedding); the pooling layer reduces the size of the feature by removing the noise and

un-needed details; the output of the previous layer is plugged to the normalization layer to normalize the input for the current layer in order to aid convergence; finally, a fully connected network is created and used for classification [18]. However, these steps are not identical for all studies. According to our investigations, most studies combined CNN with other DL algorithms such as recurrent neural network (RNN).

RNN is designed for sequence data and has the ability to remember the needed information. Therefore, it has been widely applied in sentiment analysis and sarcasm detection. The output of such networks depends on all previous computations. In other words, to predict the class of a specific word, the model may use the class of previous words and their relations. However, one of the most serious problems with this technique is gradient vanishing. To tackle this problem, Hochreiter and Schmidhuber [39] introduced LSTM and utilized it for sarcasm classification. Later, a new bidirectional version of LSTM (Bi-LSTM) was introduced. Bi-LSTM has the ability to learn from the relationships between the polar words and classify them without relying on an external lexicon. Such an approach has been found to produce better results in many studies. Another important feature is the attention layer [40], which gives the model the ability to focus on words that contribute more to sarcasm classification.

In [40], the researchers developed an attention-based Bi-LSTM model based on features learned by external pre-defined sentiment lexica, thus eliminating the need for the traditional feature vector and increasing the ability of the model to detect incongruity in sarcastic sentences. The researchers in [41] designed a hybrid system that coupled a soft attention-based Bi-LSTM with a CNN. The attention layer generates a feature vector according to which higher weights are assigned to words that are closely related to the sentence semantics. Consequently, this feature vector with pragmatic features is input in the CNN to generate the final classification. The study aimed to improve the performance in terms of accuracy, recall precision and F-measure. Another study [42] developed an attention-based Bi-LSTM model for sarcasm classification. In this model, the multi-head attention layer consists of five heads. The multiple heads allow the attention layer to move among several disjointed information spaces that reflect different representations. They used SVM for handcrafted feature extraction to be used as input for the proposed model. Another work in [43] utilized an attention-based Bi-LSTM for sarcasm classification. For better word embedding, a question answering network was designed based on five different layers, each of which provides different representations. In [44], an improved attention-based multilevel LSTM model was developed to exploit sentiment semantics in sarcasm detection. The semantic is extracted using the first-level attention-based LSTM network. Then, the sentiment semantic features obtained from the first level are used as the input for the second level. In the second level, the polarity between the sentiment semantic features and all the words in the sentence is captured to detect sarcasm by combining the LSTM and CNN networks. Later, a more complex framework was proposed in [45], in which the researchers proposed a Self-Deprecating Sarcasm (SDS) framework that incorporates GloVe embedding, CNN to extract features, bidirectional gated recurrent unit (BiGRU) to extract context information that would be useful for SDS classification, and two attention layers to assign higher weights to SDS-identified sarcastic words.

Another effective sarcasm identification system was engineered in [46] using the Bi-LSTM framework based on two main phases. In the first phase, weighted word embedding was combined with the trigram model for better word representation. In the second phase, the first phase output was inserted into a Bi-LSTM network. A novel approach was suggested in [47], in which sarcasm detection involved the sentiment of the reply to the sarcasm and the user's expression habit. In this approach, a dual-channel CNN was utilized for sarcasm detection and sentiment analysis of the reply. Moreover, attention-based LSTM was exploited to identify the user's expression habit. In a subsequent study [48], the researchers proposed a multi-head self-attention-based GRU model to detect sarcasm while considering automatic, lexical, contextual, and handcrafted features. Feature embedding was performed by a pretrained model and was enhanced using the multi-head self-attention layers to identify keywords that contribute more to classification. In [49], the researchers proposed a novel multi-task system for joint sarcasm and sentiment analysis. The local features are obtained using BiGRU, and the global features are obtained by attention-based CNN. In [50], the researchers proposed a novel feature selection approach with deep belief for detection of cyberbullying on social networks. Additionally, the Salp Swarm Algorithm was exploited to tune the network parameter for better classification accuracy. In a subsequent study [51], an attention-based LSTM sarcasm detection model was proposed to combine both hand-crafted features that are usually extracted from classical ML algorithm, such as verbs, nouns, and adjectives, with automatic features that are extracted by DL approaches. That is, the attention layer is utilized to assign weights to the words according to their level of contribution to sarcasm detection. Moreover, 16 different textual classical features are extracted and combined with the automatic features generated from the attention layer. The main contribution in this study was the proposed feature engineering approach.

To capture the variation in the performance of different classification techniques , the researchers in [52] applied five different ML algorithms, namely, Naïve Bayes, KNN, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), C4.5 Decision Tree, and SVM. Moreover, a CNN network was implemented. Additionally, different pre-processing methods were applied with the classifier to obtain the best results. In fact, a pre-trained model can be used for data preprocessing, as described in the approach in [53]. The BERT model is used for data preprocessing by converting the text into distinct tokens, and the tokens are further processed by four CNN layers. The output of this process is plugged into the LSTM layer for classification. In [10], the researchers proposed a system that combines the classical ML approach to extract different text patterns with sarcasm detection using the LSTM classifier. The basic pre-processing steps are performed on the original text before the classification. Further, in [54], the researchers proposed a new attention-based BiGRU for detecting sarcasm in which hyper parameter tuning is

performed using an artificial flora algorithm and embedding is performed by the GloVe model.

Very few works have utilized an ensemble of ML and DL approaches. One such study [55] proposed the use of a DL model in combination with an ML classifier to extract the target of sarcasm from the text. The researchers started by using an ensemble of classifiers consisting of RF, SVM, and logistic regression to classify sarcastic sentences and determine whether they contain a target. On the other hand, an LSTM is used to extract the target using aspect-based sentiment analysis.

E. Transformer-based Approaches

The sequence-to-sequence (seq2seq) model is used for many purposes, one of which is language translation, for example, translating Chinese into English [56]. One of the main disadvantages of the Seq2Seq model is that it cannot be applied to long sentences or perform parallelization. The main solution for this limitation was proposed in December 2017 in an article titled "Attention Is All You Need," which described a model called the "original transformer model" that laid the basis for transformer-based approaches [57]. In the field of NLP, a transformer can be described as a novel architecture that can solve Seq2Seq tasks while handling long-range dependencies. In addition, transformer models are trained on large-scale corpora to learn universal language representations, so the need to train a new model from scratch is eliminated [58].

Most recent studies are based on transformer models that exhibited strong performance in sarcasm detection [59], [60]. These architecture models are frequently based on transformer models such as Bidirectional Encoder Representations from Transformers (BERT) and OpenAI Generative Pre-Training-3 Model (GPT-3) [61], [62]. Recently, many researchers have been focusing on transformer models: for example, in 2021, the authors in [63] developed a context-based feature technique to detect sarcasm based on the DL model, BERT model, and conventional ML model. Two Twitter benchmark datasets, one provided by Riloff and one by Ghosh and Veale, were utilized [64], [65]; in addition, the Internet Argument Corpus (IAC-v2) benchmark was also applied. A related study [60] proposed an enhancement to BERT in order to improve its ability to handle the volume, velocity, and veracity of data.

Similarly, in 2022, the researchers in [66] introduced an enhancement to the BERT model by fine-tuning it to related intermediate tasks before applying it to the target task. The authors in [67] applied the pre-trained COMET model to generate relevant commonsense knowledge. The experiment was conducted on three datasets, including Ghosh and Ptácek from Twitter and SARC-Pol from Reddit [35], [65], and [68]. The researchers in [59] proposed a model called Contextual Response Augmentation (CRA) which uses of BERT, BiLSTM, and NetXtVLAD. The dataset consisted of Twitter and Reddit posts. To evaluate the proposed model, the IAC-V12 and AC-V23 datasets [69] and two datasets collected by Riloff et al. [64] and Ptáček et al. [35] were used. Furthermore, two datasets from Reddit [68] were utilized.

Another study in [70] developed an RCNN-RoBERTa model to tackle figurative language in social networks. This model consists of a pre-trained RoBERTa model combined with a recurrent CNN. The Semantic Evaluation Workshop Task 3 (SemEval-2018) dataset was used to measure the performance of the proposed model. Another researcher [71]proposed an encoder model called LMTweets, which is an ensemble of multiple types of techniques. Five classical classifiers, six DL algorithms, and transformer models were utilized for classification in this model. The experiments were conducted on three datasets, namely, Twitter SemEval-2018-Task, Self-Annotated Reddit Corpus (SARC), and Riloff Sarcastic Dataset [72], [64], [68].

V. Evaluation Metrics

One of the most significant aspects of most articles on models for sarcasm detection is performance evaluation, because the results provide an indication of the significance of a study. In this section, the common evaluation metrics used to assess sarcasm detection in the selected articles will be discussed. Confusion matrix is used for analyzing the performance of a binary-class model by depicting the relationship between the actual class and the predicted class. In this matrix, each row contains information about an actual class, while each column contains information about a predicted class. Accordingly, the confusion matrix aims to analyze how well a classification can recognize instances of different classes. Table III illustrates the confusion matrix [73].

TABLE III. Confusion Matrix

Class	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

In the sarcasm detection problem, true positives (TPs) are considered as sarcastic tweets that are correctly classified as sarcastic text, and true negatives (TNs) are tweets which are not sarcastic that are correctly classified as not being sarcastic (i.e., these refer to correct decisions, which are represented by the diagonal in the confusion matrix). In contrast, false positives (FPs) are instances which are not sarcastic that are misclassified as sarcastic text, and false negatives (FNs) are sarcastic tweets which are misclassified as text that is not sarcastic. The following subsection describes the most common and significant metrics for evaluation with the confusion matrix.

A. Accuracy

Accuracy is a common external measurement that reflects the percentage of the total number of tweets that are correctly classified as sarcastic or not sarcastic. It is calculated using the following equation, in which the denominator represents the total number of sarcastic tweets.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\times 100 \quad (1)$$

TABLE IV. SUMMARY OF THE EVALUATION METRICS AND PERFORMANCE OF THE REVIEWED ARTICLES

No.	Articles	Accuracy	F1-Score	Precision	Recall	AUC
1	[33]	99.1	89.0	90.0	89.0	-
2	[31]	92.7	95.5	93.0	98.3	-
3	[34]	94.0	94.0	95.0	94.0	-
4	[36]	-	69.0	-	-	-
5	[32]	77.3	69.0	69.0	69.0	-
6	[38]	-	94.9	94.7	95.2	-
7	[40]	95.3	99.0	-	-	-
8	[41]	97.9	93.5	92.1	96.8	-
9	[42]	-	77.4	72.6	83.0	-
10	[43]	-	70.8	68.9	72.8	-
11	[44]	-	87.1	85.7	89.2	-
12	[45]	93.0	94.0	92.0	98.0	-
13	[46]	95.3	-	-	-	-
14	[47]	73.0	76.0	-	-	-
15	[48]	-	98.7	97.9	99.6	99.6
16	[49]	92.2	-	91.6	92.0	-
17	[50]	99.0	94.0	-	-	-
18	[51]	-	99.0	99.0	99.0	-
19	[52]	-	66.0	-	-	-
20	[53]	99.6	99.5	99.3	99.8	-
21	[10]	-	82.3	89.3	76.4	72.2
22	[54]	96.8	97.0	97.2	97.2	-
23	[55]	21.7	54.9	-	-	-
24	[59]	-	93.1	93.2	93.6	-
25	[60]	70.6	70.5	68.7	72.5	-
26	[63]	99.0	99.0	98.0	99.5	-
27	[66]	-	97.4	-	-	-
28	[67]	-	85.4	85.7	86.1	-
29	[70]	91.0	90.0	90.0	90.0	94.0
30	[71]	75.0	74.0	73.0	85.0	76.0

B. F1-Score

F1-score is a combination of precision and recall measures, which are the most frequently used metrics. Indeed, to calculate F1-score, precision and recall need to be calculated using the equations (2) and (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100 \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100 \quad (3)$$

As mentioned before, F1-score is calculated as a harmonic mean of precision and recall, as demonstrated in the equation below.

$$F(i,j) = \frac{(2 \times Precision \times Recall)}{Precision + Recall} \times 100 \quad (4)$$

In general, F1-score values are within the interval [0, 1]; therefore, the higher the F1-score value, the better is the classification. Table IV presents a summary and comparison of the evaluation metrics used in sarcasm detection in the selected articles. The table shows that more than four types of evaluation metrics have been applied to evaluate sarcasm detection. From the table, it can be observed that the most common measures are F1-score, precision, and recall, and they are followed by accuracy. The results in this table are based on the highest results reported by studies that used multiple algorithms or multiple datasets.

VI. DATASET COLLECTION

Dataset collection is a crucial step in the sarcasm classification process that can affect the entire procedure. Building and annotating datasets for sarcasm detection is a challenging task even for human annotators, since the sarcastic text could be implicit, ambiguous, and hard to identify [74], [75]. It is normal for disagreements between annotators regarding the classification of a single text as sarcastic or not, so the task is even harder for an AI program. This section describes the datasets that were used in the reviewed literature. Noticeably, some articles utilized datasets that were used in the reviewed literature. Also, some articles utilized datasets from multiple sources, including social networks, news headlines, sarcastic reviews on online shops, books snippets, and forums, to stress on the generalization of their systems. For instance, the researchers in [55] utilized three different datasets, including book snippets, tweets, and Reddit comments. However, other articles relied on a single source for the dataset, for example, social network posts [52].

Social network posts have limited length; for example, Twitter limits tweets to 280 characters. This makes it simpler to obtain annotated text based on hashtags and API. The monthly number of active users on Twitter is about 330 million, which makes it a rich source of sarcastic tweets [11]. Therefore, most of the reviewed studies rely on Twitter as a source for their datasets, and few articles used datasets from Reddit and other sources. Twitter-based datasets can be built automatically using the Twitter streaming API while searching for a specific hashtag, such as "#Irony" and "#sarcasm" [32]. In this case, the annotation process is guided by the hashtag itself. In addition, it is already accurate to some extent, since the author clearly declares the sarcasm in the tweet. Another process for collecting datasets is manual self-annotation. For instance, in [60], the annotation process was undertaken by three linguistic annotators, each of whom worked on a subset of the dataset, and in [32], four expert annotators participated in the dataset annotation. In [52], the annotation was manually performed by three students. To ensure the reliability of the self-annotation, an evaluation step can be performed later by a third annotator [60].

The number of the collected instances of sarcasm in the considered datasets varied from 1264 to 1055277 [49], [76]. Generally, the higher the number of tweets in the dataset, the higher is the effectiveness of the proposed models. Some studies used a public dataset, such as [44], while other articles collected data on their own, such as [10]. When a public dataset is used for evaluation, it allows for fairer and more meaningful comparison with other works that use the same dataset.

The number of sarcastic and non-sarcastic samples in the dataset obviously affect the performance of the detection model. An imbalanced dataset may skew the performance of the classification model. In general, the models developed using imbalanced datasets are likely to achieve greater accuracy than other models with conflicting F1-score values [77]. For example, the Riloff dataset [64] creates a bias toward non-sarcastic tweets as it consists of 1648 non-sarcastic tweets and 308 sarcastic tweets. A detailed description of the datasets in the reviewed articles is presented in Table V.

TABLE V. DESCRIPTION OF THE DATASETS USED IN THE REVIEWED ARTICLES

No.	Article	Dataset source	Accessibility	#Instances	Annotation	#Sources	Balance
1	[33]	Twitter	Private	NA	Hashtag	Single	NA
2	[31]	Twitter	Private	76,799	Both	Single	NA
3	[34]	Twitter	Public	780,000	Hashtag	Single	No
4	[36]	Twitter	Private	5000	NA	Single	Yes
5	[32]	Twitter	Private	6600	Both	Single	Yes
6	[38]	Others	Public	28501	Self-annotated	Single	Yes
7	[40]	7 datasets from Twitter	Public	12162	Both	Multiple	4 yes, 3 no
8	[41]	Twitter1, Twitter2	Public/Private	55961	Self-annotated	Multiple	Yes, no
9	[42]	Reddit	Public	6534	Self-annotated	Single	Yes
10	[43]	Others	Public	4692	Self-annotated	Single	Yes
11	[44]	Others, others, Twitter	Public	55795	NA	Multiple	NA
12	[45]	7 datasets from Twitter	Public	134407	Both	Multiple	Yes
13	[46]	Twitter, others, others	Public	40000	Self-annotated	Multiple	No, yes
14	[47]	Reddit, Twitter	Public	45301	Both	Multiple	NA
15	[48]	Twitter1, others, Twitter2, Reddit, others	Public	309566	Both	Multiple	4 yes, 1 no
16	[49]	Others, Twitter	Public/Private	1264	Hashtag	Multiple	No, yes
17	[50]	NA	Private	NA	NA	Single	NA
18	[51]	Twitter1, Twitter2, others	Public	83596	Both	Multiple	2 yes, 1 no
19	[52]	Twitter	Public	4618	Self-annotated	Single	Yes
20	[53]	Others, others	Public	55328	Self-annotated	Multiple	Yes
21	[10]	Tweets, Reddit, Others	Private	20000	Self-annotated	Multiple	Yes
22	[54]	Others	Public	28,501	Self-annotated	Single	Yes
23	[55]	Twitter, Reddit, others	Public	1680	Self-annotated	Multiple	NA
24	[59]	Twitter, Reddit	Private	13000	NA	Multiple	NA
25	[60]	Twitter	Public	3000	Self-annotated	Single	NA
26	[63]	Twitter1, Twitter2, others	Public	58436	Hashtag	Multiple	No, yes, yes
27	[66]	Others, Reddit, Twitter	Public	1018291	Self-annotated	Multiple	Yes, yes, no
28	[67]	Twitter1, Twitter2, Reddit	Public	65551	Self-annotated	Multiple	NA
29	[70]	Twitter, Twitter, Reddit, Twitter	Public	NA	NA	Multiple	NA
30	[71]	Twitter, Reddit	Public	47115	Self-annotated	Multiple	No

VII. DISCUSSION

This SLR analyzed 30 articles that were able to address the four research questions. This section discusses the findings of the review, highlights the challenges, and provides future research directions that can help in the development of more accurate and efficient sarcasm detection tools.

A. Findings

In several domains, NLP is an increasingly important topic with regard to AI and its applications. The research community is paying close attention to the sarcasm detection approaches, datasets and metrics. This subsection focuses on several observations from examination of different aspects of sarcasm detection.

1) Approaches: In general, it is impossible to compare the different approaches objectively due to several variations in the dataset sources and task requirements. One of the most interesting findings, as shown in Fig. 3, is that more than half of the reviewed articles used DL as a classification method for sarcasm detection, and there was a noticeable upward trend in the application of DL techniques for solving several NLP problems. In fact, DL has proved its superiority in sentiment analysis, in general, and in sarcasm detection in particular. One possible reason for this is that the automated feature extraction aspect is more effective and gives better insights about the target text than handcrafted features used in other classical sarcasm detection techniques. There are, however, other possible explanations. For instance, with regard to model performance, it is found that the best accuracy for the reviewed articles was obtained with DL models. Moreover, specific DL techniques, such as RNN, are particularly designed for sequence input data, and this fits the requirements of sarcasm detection tasks.

In addition, as depicted in Fig. 3, an interesting observation was that most articles used hybrid approaches in order to exploit the advantages of more than one approaches. The hybrid approach is extremely important in the development of sarcasm detection tools, as demonstrated in several articles in Section IV. Moreover, classical ML algorithms were utilized by 16% of the researchers. In contrast, only a few of the reviewed articles utilized transformer-based approaches. This is probably because transformers are a relatively new invention for application to sarcasm detection models. However, the rapid improvement in computational resources and increase in the available datasets have led to an increase in the application of transformer-based approaches in recent times.

Fig. 4 depicts the frequency at which various sarcasm detection techniques were used in the reviewed articles in this SLR. Among the classical ML approaches, the most commonly used classifier is SVM. Moreover, for DL approaches, the most commonly used technique is Bi-LSTM, and for transformers, the most applied technique is BERT. To sum up, the most frequently utilized sarcasm detection approach is DL. Moreover, the transformer approach appears to be an emerging promising solution with comparative performance to currently popular techniques and it warrants further investigation.

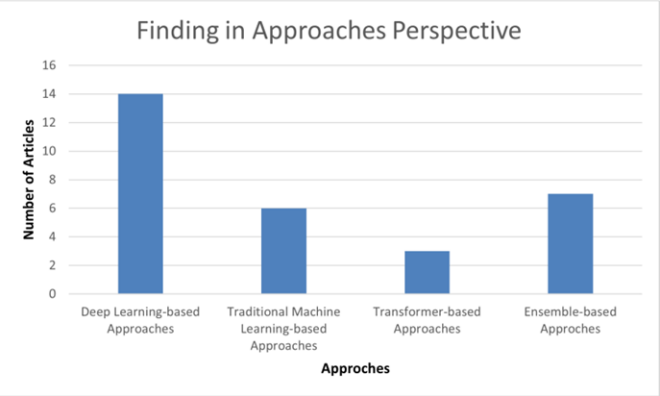

Fig. 3. Trends in the sarcasm detection approaches used by the reviewed articles.

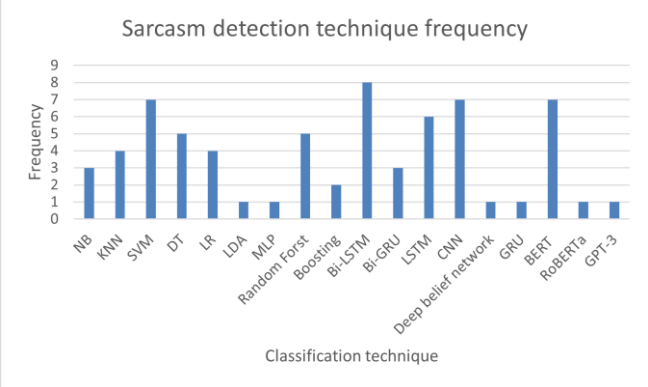

Fig. 4. Classification techniques used for sarcasm detection in the reviewed articles.

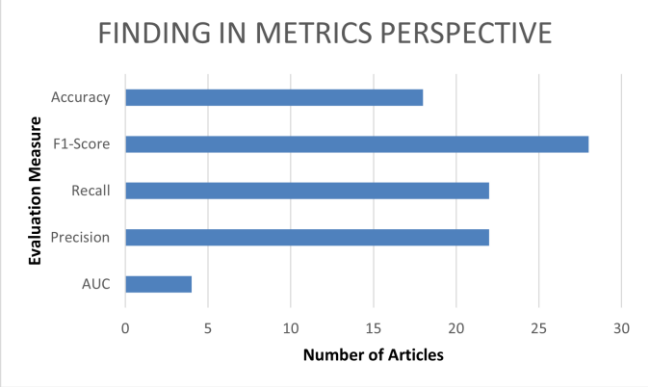

Fig. 5. Frequency of the use of various metrics for evaluation of sarcasm detection models in the included articles.

2) Metrics: As discussed in Section V, researchers used precision, accuracy, recall, F1-score, and AUC as evaluation metrics. As shown in Fig. 5, one of the most significant findings from this SLR is that the majority of researchers utilized F-score, followed by precision and recall. Furthermore, the most obvious finding to emerge from the analysis is that 10% of the reviewed articles used AUC as the evaluation metric. In addition, from the data in Fig. 5, it is apparent that accuracy was used as a metric by 63% of the researchers.

Overall, none of the evaluation metrics fit all sarcasm detection problems due to differences in the characteristics of datasets and approaches used. It is not surprising that F1-score was the most frequently used metric (90% of the researchers used this metric). This is probably because the F-score can balance the precision and recall of the positive class. Moreover, the F1-score could be more suitable than other measures when the target classes are unevenly distributed. Another interesting observation was the correlation between accuracy and dataset balance in the reviewed articles, since the vast majority of datasets were balanced datasets. This may explain why the use of accuracy as an evaluation metric was as high as 63% in the reviewed articles. AUC was the least frequently used metric; this is probably because AUC is based only on the thresholds of the true positive rate and the false positive rate. This is in contrast to the F1-socre, which takes into account the overall recall and precision values. In general, 87% of the observed studies used more than two metrics, and this makes the evaluation framework more robust.

3) Datasets: The dataset sources, number of datasets, dataset accessibility, number of instances, annotation methods, and dataset types of the included articles are discussed here.

An essential factor that affects the sarcasm detection process is the source of the dataset, as shown in Fig. 6. The findings showed that 34% of the analyzed articles used Twitter as a unique source of datasets. One possible reason for this is the huge number of Twitter users, which is 330 million monthly active users [11]. Moreover, Twitter provides concise text that can be automatically annotated by hashtags, and this facilitates dataset building. However, no single public dataset was used across all the reviewed articles.

The most obvious finding to emerge from the analysis is that 50% of the reviewed articles rely on heterogeneous dataset sources. This result may be explained by the different advantages offered by different sources. For instance, Twitter provides short texts while Facebook provides longer texts. Therefore, considering different sources for model building is expected to produce a more comprehensive classification model. Fig. 7 supports this notion, as it shows that 63% of the reviewed studies used multiple datasets rather than a single dataset.

Another important finding that strongly supports the transparency of the evaluation framework is that 71% of the considered articles used public datasets, 23% used private datasets, and 6% used both private and public datasets, see Fig. 8. This enabled the researchers to conduct a fair comparison of the proposed work with others conducted with the same public dataset. Additionally, 73% of the reviewed articles used less than 100,000 instances to build their classification model, while only 17% used more than 100,000 instances, as shown in Fig. 9. A possible explanation for this is that sarcasm detection tasks do not require a huge dataset to differentiate between sarcastic and non-sarcastic text. This is supported by the finding that good performance was observed for most datasets containing less than 100,000 instances. Moreover, the computation overhead is a serious concern when it comes to building a classification model.

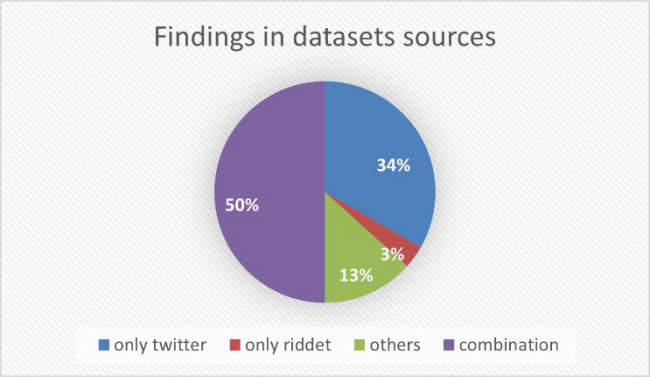

Fig. 6. Dataset sources of the included articles.

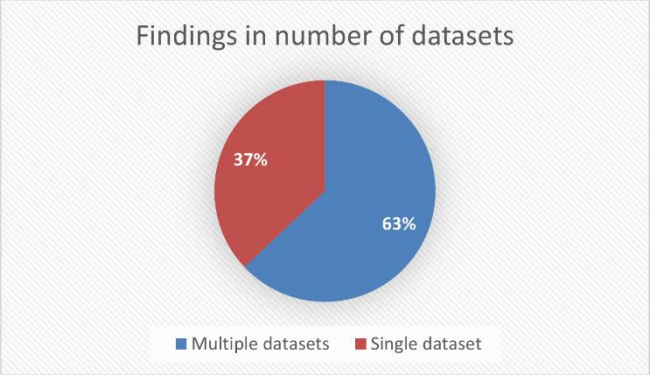

Fig. 7. Use of single or multiple datasets in the reviewed articles.

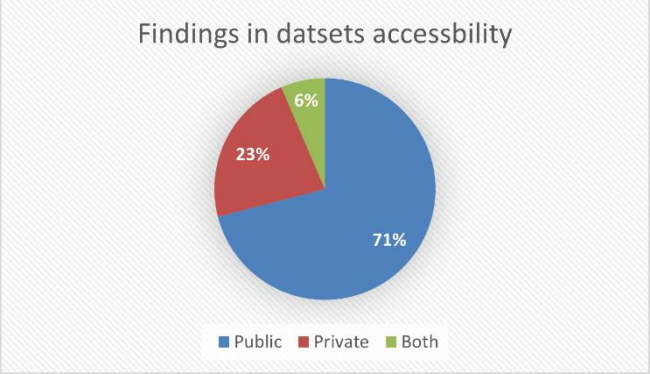

Fig. 8. Accessibility of datasets used in the reviewed articles.

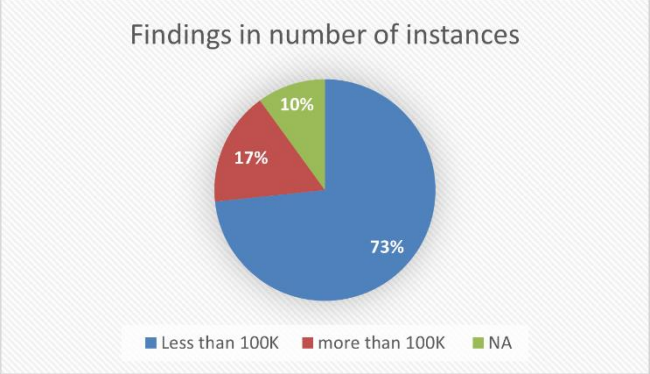

Fig. 9. Number of instances evaluated in the reviewed articles.

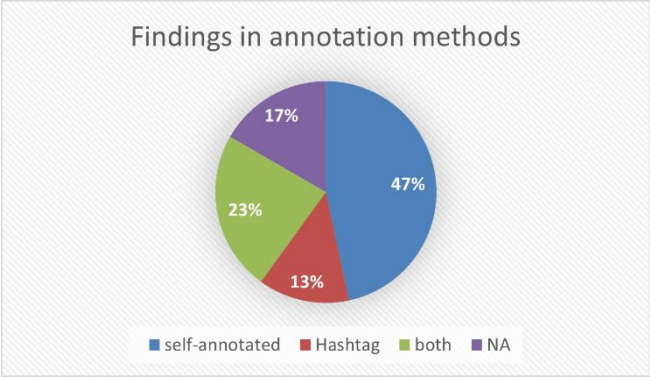

Fig. 10. Annotation methods used in the reviewed articles.

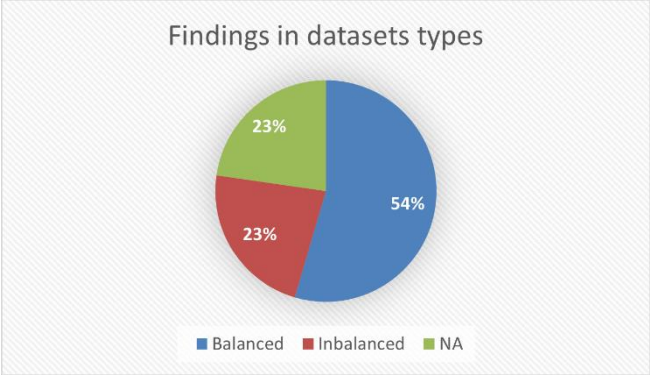

Fig. 11. Distribution of dataset types in the reviewed articles.

Another important issue related to datasets is the annotation method. As expected, 47% of the analyzed articles used self-annotated datasets, illustrated in Fig. 10. Self-annotated datasets are precise because the text is analyzed and annotated by experts and reviewed by another group of experts. However, self-annotation requires a tremendous amount of time [13]. Therefore, tweets could be annotated automatically based on the hashtag included in the tweet; this is a simple and time-conserving approach for annotations that has an acceptable level of correctness. However, only 13% of the considered articles used hashtag-based annotation, and 23% used both the self-annotation and hashtag annotation methods.

Another relevant finding was that 54% of the used datasets were balanced datasets in which the number of sarcastic and non-sarcastic instances was similar, as shown in Fig. 11. This is probably because the nature of the dataset highly influences the model prediction metrics, particularly accuracy and F-measure. These findings reflect those of Eke et al. [13], who also found that an imbalanced dataset can increase the accuracy of the model.

B. Open Research Questions

This subsection discusses the common issues and main challenges in the development of sarcasm detection tools for social networks, based on the findings from the reviewed articles.

1) Language used in the social network: The language used in social networks is not only restricted with regard to grammar, but also restricted to words that are not often included in dictionaries. This might pose an additional challenge in the recognition of sarcasm on Twitter and Reddit because of typos, non-vocabulary language, and non-grammatical context. As multilingual text has recently grabbed the attention of researchers, training models in more than one language might be more efficient.

2) Dataset: One of the biggest challenges in training models is the skewness of data. This problem arises when the number of instances in one class, such as sarcastic text, is greater than that in the other class, that is, non-sarcastic text. Furthermore, the quality of the dataset is another challenge. The use of a mixed dataset that uses slang and informal language makes it more difficult to train the classification model, especially if the dataset does not contain hashtags. In such a scenario, creating standard datasets is a solution that may solve the mentioned problems.

3) Text-based sarcasm detection: In speech, sarcasm detection includes features such as eye contact and body language, which help in the recognition of sarcasm. However, text data lack such features. Therefore, it is difficult and takes considerably more effort to identify sarcasm in text.

4) Variable context length: According to the reviewed articles, finding the optimal length of conversational context is a challenge. The Twitter dataset is the most commonly used domain for sarcasm detection, but the short text can be noisy and may not have any relevant features. Therefore, detecting sarcasm from short text is difficult. Overall, the researchers' task is still challenging due to the variability in context length.

5) Emoticons and special characters: In the last decade, the use of emoticons and special characters in social networks has increased. Most people prefer to express their feeling through emojis and emoticons, especially in applications that have restrictions on the number of characters such as Twitter. This increases the likelihood of ambiguity and makes sarcasm detection more difficult. Therefore, researchers should take into account the importance of these features, as they may change the overall sentiment of the sentence.

6) Data annotation: The manual annotation method is a major challenge. The main problem is distinguishing between perceived and intended sarcasm. Most datasets built through manual annotation may, therefore, be limited by differences in the perception of the annotator and the intention of the author of the utterance. As the labeling is based on the perceived sarcasm, this may lead to false positives and false negatives. A solution for this was proposed in [78], according to which the annotator and author of the utterances should be the same individual. Moreover, manual annotation requires a lot of time and the recruitment of domain experts.

7) Lack of real-time sarcasm detection: With the increase in the volume of generated data on social networks, sarcasm detection in real time is a challenging but significant task. Despite this, none of the reviewed articles included real-time data analysis.

Overall, there are still several challenges and open problems in sarcasm detection that need to be worked on. The following subsection provides future research directions.

C. Future Research Directions

This section describes the possible research directions based on our analysis of the 30 articles.

1) Considering more languages: The majority of the recent sarcasm detection works focus on English and ignore other languages. To this end, one possible future direction is to consider multiple language models that have the ability to perform all sarcasm detection sub-tasks for multiple languages.

2) Application of transformers and DL models: While considerably more work will need to be done on transformer-based, DL-based, and hybrid systems, their performance is superior to that of ML and classical NLP techniques. Moreover, the amount of work on transformer-based approaches is still limited, and therefore, there is scope for the development of more transformer-based sarcasm detection models.

3) Tweet correctness techniques: The findings in the datasets indicates that Twitter is the most frequently used source of data for sarcasm detection model evaluation in the reviewed articles. However, tweets are likely to have many typos, which may negatively influence model performance. One possible future direction is to use an automatic technique for typo correction in the early stage of development of sarcasm detection systems.

4) Exploring other social network sources: Twitter and Reddit were the only dataset sources in the reviewed articles. While they are both good sources of data, the addition of more social networks sources would provide a more comprehensive model. Therefore, further work in this domain should focus more on other social networks sources such as Facebook and Instagram.

5) Multi-culture datasets: Sarcasm by its nature differs across cultures. In fact, there could be cultural differences even between people who speak the same language. Therefore, further research could focus on the relationships between culture and sarcasm and the detection of sarcasm in multi-culture datasets.

6) Building multimodal sarcasm detection models: Most of the recent work on sarcasm detection focuses only on text-based datasets. However, considering multimodal models is a good idea for exploring new methods to solve such problems.

7) Use of emojis and emotions: Sarcastic text on social network often contains emojis that are used to express a specific emotion, due to the limitations on the length of posts on some platforms. Therefore, more research is required on new ideas for dealing with data that can improve the performance of such classification models.

VIII. Conclusions

Recently, sarcasm detection, especially in social networks, has grabbed the attention of many researchers. This SLR covers articles on sarcasm detection to answer four research questions. The review of the selected studies provides an analysis of the current approaches, metrics and datasets used to evaluate their models, as well as the challenges facing the development of sarcasm detection applications. In this SLRs, 30 articles published between 2019 and 2022 obtained from four well-known digital databases in Computer Science were analyzed based on their approaches, datasets, and evaluation metrics. Moreover, challenges and open research problems that still prevail in sarcasm detection are discussed. The findings show that the DL approach is most widely utilized, and it is followed by hybrid approaches. Furthermore, Twitter is the most commonly utilized source for datasets, and most researchers used public heterogeneous datasets. With regard to the features of the datasets, most studies used balanced datasets, and there is no consensus among researchers about whether standard, publicly available datasets are suitable for sarcasm detection in social networks. With regard to performance metrics, precision, recall, accuracy, and F1-score were most frequently used in the selected articles, and the majority of the articles used F1-score. Finally, several recommendations, including considering more languages, building multimodal sarcasm detection models and tweet correctness techniques have been suggested to improve the efficiency and performance of sarcasm detection tools.

References

[1] B. Liu and L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis,” in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, 2012, pp. 415–463. doi: 10.1007/978-1-4614-3223-4_13.

[2] “Biggest social media platforms 2022,” Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed Oct. 27, 2022).

[3] B. Yee Liau and P. Pei Tan, “Gaining customer knowledge in low cost airlines through text mining,” Ind. Manag. Data Syst., vol. 114, no. 9, pp. 1344–1359, Jan. 2014, doi: 10.1108/IMDS-07-2014-0225.

[4] 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, Aug. 2016, doi: 10.1016/j.dcan.2016.06.002.

[5] S. G. Wicana, T. Y. Ibisoglu, and U. Yavanoglu, “A Review on Sarcasm Detection from Machine-Learning Perspective,” in 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 2017, pp. 469–476. doi: 10.1109/ICSC.2017.74.

[6] U. Yavanoglu, T. Y. Ibisoglu, and S. G. Wıcana, “Sarcasm Detection Algorithms,” Int. J. Semantic Comput., vol. 12, no. 03, pp. 457–478, Sep. 2018, doi: 10.1142/S1793351X18300017.

[7] Y. Kumar and N. Goel, “AI-Based Learning Techniques for Sarcasm Detection of Social Media Tweets: State-of-the-Art Survey,” SN Comput. Sci., vol. 1, no. 6, p. 318, Nov. 2020, doi: 10.1007/s42979-020-00336-3.

[8] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, “Sarcasm detection using machine learning algorithms in Twitter: A systematic review,” Int. J. Mark. Res., vol. 62, no. 5, pp. 578–598, Sep. 2020, doi: 10.1177/1470785320921779.

[9] F. B. Kader, N. H. Nujat, T. B. Sogir, M. Kabir, H. Mahmud, and K. Hasan, “Computational Sarcasm Analysis on Social Media: A Systematic Review.” arXiv, Sep. 20, 2022. Accessed: Oct. 29, 2022. [Online]. Available: http://arxiv.org/abs/2209.06170.

[10] M. Bouazizi and T. Ohtsuki, “Sarcasm Over Time and Across Platforms: Does the Way We Express Sarcasm Change?,” IEEE Access, vol. 10, pp. 55958–55987, 2022, doi: 10.1109/ACCESS.2022.3174862.

[11] A.-C. Băroiu and Ștefan Trăușan-Matu, “Automatic Sarcasm Detection: Systematic Literature Review,” Information, vol. 13, no. 8, p. 399, Aug. 2022, doi: 10.3390/info13080399.

[12] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," vol. 2, Jan. 2007.

[13] C. I. Eke, A. A. Norman, Liyana Shuib, and H. F. Nweke, "Sarcasm identification in textual data: systematic review, research challenges and open directions," Artif. Intell. Rev., vol. 53, no. 6, pp. 4215–4258, Aug. 2020, doi: 10.1007/s10462-019-09791-8.

[14] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," J. Syst. Softw., vol. 125, pp. 207–219, Mar. 2017, doi: 10.1016/j.jss.2016.11.027.

[15] D. G. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," in Lrec 2014 proceedings, 2014.

[16] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in Twitter data," in 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2015, pp. 1373–1380. doi: 10.1145/2808797.2808910.

[17] T. Veale and Y. Hao, "Detecting Ironic Intent in Creative Comparisons," presented at the ECAI 2010, 2010, pp. 765–770.

[18] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowl.-Based Syst., vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[19] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[20] V. Hatzivassiloglou and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives," in 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, Jul. 1997, pp. 174–181. doi: 10.3115/976909.979640.

[21] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," ACM Trans. Inf. Syst., vol. 21, no. 4, pp. 315–346, Oct. 2003, doi: 10.1145/944012.944013.

[22] H. Han, J. Zhang, J. Yang, Y. Shen, and Y. Zhang, "Generate domain-specific sentiment lexicon for review sentiment analysis," Multimed. Tools Appl., vol. 77, no. 16, pp. 21265–21280, Aug. 2018, doi: 10.1007/s11042-017-5529-5.

[23] Q. Cao, W. Duan, and Q. Gan, "Exploring determinants of voting for the 'helpfulness' of online user reviews: A text mining approach," Decis. Support Syst., vol. 50, no. 2, pp. 511–521, Jan. 2011, doi: 10.1016/j.dss.2010.11.009.

[24] O. Araque, G. Zhu, and C. A. Iglesias, "A semantic similarity-based perspective of affect lexicons for sentiment analysis," Knowl.-Based Syst., vol. 165, pp. 346–359, Feb. 2019, doi: 10.1016/j.knosys.2018.12.005.

[25] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining," Procedia Comput. Sci., vol. 46, pp. 635–643, Jan. 2015, doi: 10.1016/j.procs.2015.02.112.

[26] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "A semantic matching energy function for learning with multi-relational data," Mach. Learn., vol. 94, no. 2, pp. 233–259, Feb. 2014, doi: 10.1007/s10994-013-5363-6.

[27] W. Zhang, H. Xu, and W. Wan, "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," Expert Syst. Appl., vol. 39, no. 11, pp. 10283–10291, Sep. 2012, doi: 10.1016/j.eswa.2012.02.166.

[28] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Reviewing Classification Approaches in Sentiment Analysis," in Soft Computing in Data Science, Singapore, 2015, pp. 43–53. doi: 10.1007/978-981-287-936-3_5.

[29] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May 2010. Accessed: Oct. 14, 2022. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[30] S. Park and Y. Kim, "Building thesaurus lexicon using dictionary-based approach for sentiment classification," in 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Jun. 2016, pp. 39–44. doi: 10.1109/SERA.2016.7516126.

[31] K. Sundararajan and A. Palanisamy, "Multi-Rule Based Ensemble Feature Selection Model for Sarcasm Type Detection in Twitter," Comput. Intell. Neurosci., vol. 2020, pp. 1–17, Jan. 2020, doi: 10.1155/2020/2860479.

[32] V. Govindan and V. Balakrishnan, "A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, pp. 5110–5120, Sep. 2022, doi: 10.1016/j.jksuci.2022.01.008.

[33] J. Godara, I. Batra, R. Aron, and M. Shabaz, "Ensemble Classification Approach for Sarcasm Detection," Behav. Neurol., vol. 2021, pp. 1–13, Nov. 2021, doi: 10.1155/2021/9731519.

[34] 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, doi: 10.1109/ACCESS.2021.3076789.

[35] T. Ptáček, I. Habernal, and J. Hong, "Sarcasm Detection on Czech and English Twitter," in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, Aug. 2014, pp. 213–223. Accessed: Oct. 29, 2022. [Online]. Available: https://aclanthology.org/C14-1022.

[36] A. Khatri, P. P, and D. A. K. M, "Sarcasm Detection in Tweets with BERT and GloVe Embeddings." arXiv, Jun. 20, 2020. Accessed: Oct. 29, 2022. [Online]. Available: http://arxiv.org/abs/2006.11512.

[37] M. Choli and D. J. Kuss, "Perceptions of blame on social media during the coronavirus pandemic," Comput. Hum. Behav., vol. 124, p. 106895, Nov. 2021, doi: 10.1016/j.chb.2021.106895.

[38] D. Vinoth and P. Prabhavathy, "An intelligent machine learning-based sarcasm detection and classification model on social networks," J. Supercomput., vol. 78, no. 8, pp. 10575–10594, May 2022, doi: 10.1007/s11227-022-04312-x.

[39] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[40] S. Zhang, X. Zhang, J. Chan, and P. Rosso, "Irony detection via sentiment-based transfer learning," Inf. Process. Manag., vol. 56, no. 5, pp. 1633–1644, Sep. 2019, doi: 10.1016/j.ipm.2019.04.006.

[41] 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, doi: 10.1109/ACCESS.2019.2899260.

[42] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti, "Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM," IEEE Access, vol. 8, pp. 6388–6397, 2020, doi: 10.1109/ACCESS.2019.2963630.

[43] Y. Diao et al., "A Multi-Dimension Question Answering Network for Sarcasm Detection," IEEE Access, vol. 8, pp. 135152–135161, 2020, doi: 10.1109/ACCESS.2020.2967095.

[44] 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, doi: 10.1016/j.neucom.2020.03.081.

[45] A. Kamal and M. Abulaish, "CAT-BiGRU: Convolution and Attention with Bi-Directional Gated Recurrent Unit for Self-Deprecating Sarcasm Detection," Cogn. Comput., vol. 14, no. 1, pp. 91–109, Jan. 2022, doi: 10.1007/s12559-021-09821-0.

[46] A. Onan and M. A. Toçoğlu, "A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification," IEEE Access, vol. 9, pp. 7701–7722, 2021, doi: 10.1109/ACCESS.2021.3049734.

[47] 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," Cogn. Comput., vol. 14, no. 1, pp. 78–90, Jan. 2022, doi: 10.1007/s12559-021-09832-x.

[48] R. Akula and I. Garibay, "Interpretable Multi-Head Self-Attention Architecture for Sarcasm Detection in Social Media," Entropy, vol. 23, no. 4, Art. no. 4, Apr. 2021, doi: 10.3390/e23040394.

[49] Chunyan Yin, Y. Chen, and W. Zuo, "Multi-Task Deep Neural Networks for Joint Sarcasm Detection and Sentiment Analysis," Pattern Recognit. Image Anal., vol. 31, no. 1, pp. 103–108, Jan. 2021, doi: 10.1134/S105466182101017X.

[50] N. S et al., "Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media," Comput. Intell. Neurosci., vol. 2022, p. e2163458, Jun. 2022, doi: 10.1155/2022/2163458.

[51] R. Pandey, A. Kumar, J. P. Singh, and S. Tripathi, "Hybrid attention-based Long Short-Term Memory network for sarcasm identification," Appl. Soft Comput., vol. 106, p. 107348, Jul. 2021, doi: 10.1016/j.asoc.2021.107348.

[52] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection," Inf. Process. Manag., vol. 58, no. 4, p. 102600, Jul. 2021, doi: 10.1016/j.ipm.2021.102600.

[53] S. Bhardwaj and M. R. Prusty, "BERT Pre-processed Deep Learning Model for Sarcasm Detection," Natl. Acad. Sci. Lett., vol. 45, no. 2, pp. 203–208, Apr. 2022, doi: 10.1007/s40009-022-01108-8.

[54] "Automated sarcasm detection and classification using hyperparameter tuned deep learning model for social networks - Vinoth - Expert Systems - Wiley Online Library." https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13107 (accessed Oct. 14, 2022).

[55] P. Parameswaran, A. Trotman, V. Liesaputra, and D. Eyers, "Detecting the target of sarcasm is hard: Really??," Inf. Process. Manag., vol. 58, no. 4, p. 102599, Jul. 2021, doi: 10.1016/j.ipm.2021.102599.

[56] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in Advances in Neural Information Processing Systems, 2014, vol. 27. Accessed: Oct. 14, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

[57] A. Vaswani et al., "Attention Is All You Need." arXiv, Dec. 05, 2017. doi: 10.48550/arXiv.1706.03762.

[58] D. Jurafsky and J. Martin, Speech and Language Processing. 2022.

[59] H. Lee, Y. Yu, and G. Kim, "Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context." arXiv, Jun. 11, 2020. Accessed: Oct. 29, 2022. [Online]. Available: http://arxiv.org/abs/2006.06259.

[60] M. Shrivastava and S. Kumar, "A pragmatic and intelligent model for sarcasm detection in social media text," Technol. Soc., vol. 64, p. 101489, Feb. 2021, doi: 10.1016/j.techsoc.2020.101489.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Accessed: Oct. 29, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805.

[62] T. B. Brown et al., "Language Models are Few-Shot Learners." arXiv, Jul. 22, 2020. Accessed: Jan. 28, 2023. [Online]. Available: http://arxiv.org/abs/2005.14165.

[63] C. I. Eke, A. A. Norman, and L. Shuib, "Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model," IEEE Access, vol. 9, pp. 48501–48518, 2021, doi: 10.1109/ACCESS.2021.3068323.

[64] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as Contrast between a Positive Sentiment and Negative Situation," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, Oct. 2013, pp. 704–714. Accessed: Oct. 14, 2022. [Online]. Available: https://aclanthology.org/D13-1066.

[65] A. Ghosh and Dr. T. Veale, "Fracking Sarcasm using Neural Network," in Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, San Diego, California, 2016, pp. 161–169. doi: 10.18653/v1/W16-0425.

[66] E. Savini and C. Caragea, "Intermediate-Task Transfer Learning with BERT for Sarcasm Detection," Mathematics, vol. 10, no. 5, p. 844, Mar. 2022, doi: 10.3390/math10050844.

[67] J. Li, H. Pan, Z. Lin, P. Fu, and W. Wang, "Sarcasm Detection with Commonsense Knowledge," IEEEACM Trans. Audio Speech Lang. Process., vol. 29, pp. 3192–3201, 2021, doi: 10.1109/TASLP.2021.3120601.

[68] M. Khodak, N. Saunshi, and K. Vodrahalli, "A Large Self-Annotated Corpus for Sarcasm." arXiv, Mar. 22, 2018. Accessed: Oct. 29, 2022. [Online]. Available: http://arxiv.org/abs/1704.05579.

[69] S. Lukin and M. Walker, "Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue." arXiv, Aug. 28, 2017. Accessed: Oct. 29, 2022. [Online]. Available: http://arxiv.org/abs/1708.08572.

[70] R. A. Potamias, G. Siolas, and A.-G. Stafylopatis, "A Transformer-based approach to Irony and Sarcasm detection," Neural Comput. Appl., vol. 32, no. 23, pp. 17309–17320, Dec. 2020, doi: 10.1007/s00521-020-05102-3.

[71] R. Ahuja and S. C. Sharma, "Transformer-Based Word Embedding With CNN Model to Detect Sarcasm and Irony," Arab. J. Sci. Eng., vol. 47, no. 8, pp. 9379–9392, Aug. 2022, doi: 10.1007/s13369-021-06193-3.

[72] C. Van Hee, E. Lefever, and V. Hoste, "Exploring the fine-grained analysis and automatic detection of irony on Twitter," Lang. Resour. Eval., vol. 52, no. 3, pp. 707–731, Sep. 2018, doi: 10.1007/s10579-018-9414-2.

[73] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," J. Microbiol. Methods, vol. 43, no. 1, pp. 3–31, Dec. 2000, doi: 10.1016/S0167-7012(00)00201-3.

[74] D. Rao and D. Ravichandran, "Semi-Supervised Polarity Lexicon Induction," in Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, Mar. 2009, pp. 675–682. Accessed: Oct. 14, 2022. [Online]. Available: https://aclanthology.org/E09-1077.

[75] E. Filatova, "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing," p. 7.

[76] P. Goel, R. Jain, A. Nayyar, S. Singhal, and M. Srivastava, "Sarcasm detection using deep learning and ensemble learning," Multimed. Tools Appl., May 2022, doi: 10.1007/s11042-022-12930-z.

[77] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm Detection on Twitter: A Behavioral Modeling Approach," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai China, Feb. 2015, pp. 97–106. doi: 10.1145/2684822.2685316.

[78] S. Oprea and W. Magdy, "iSarcasm: A Dataset of Intended Sarcasm." arXiv, May 01, 2020. Accessed: Oct. 22, 2022. [Online]. Available: http://arxiv.org/abs/1911.03123.