information

MDPI

Review

Automatic Sarcasm Detection: Systematic Literature Review

Alexandru-Costin Băroiu [1] and Ștefan Trăușan-Matu [1,2,*]

[1] Faculty of Automatic Control and Computer Science, Politehnica University of Bucharest, 060042 Bucharest, Romania
[2] Research Institute for Artificial Intelligence "Mihai Draganescu" of the Romanian Academy, 050711 Bucharest, Romania
* Correspondence: stefan.trausan@upb.ro or trausan@gmail.com

Abstract: Sarcasm is an integral part of human language and culture. Naturally, it has garnered great interest from researchers from varied fields of study, including Artificial Intelligence, especially Natural Language Processing. Automatic sarcasm detection has become an increasingly popular topic in the past decade. The research conducted in this paper presents, through a systematic literature review, the evolution of the automatic sarcasm detection task from its inception in 2010 to the present day. No such work has been conducted thus far and it is essential to establish the progress that researchers have made when tackling this task and, moving forward, what the trends are. This study finds that multi-modal approaches and transformer-based architectures have become increasingly popular in recent years. Additionally, this paper presents a critique of the work carried out so far and proposes future directions of research in the field.

Keywords: automatic sarcasm detection; natural language processing; systematic literature review

Citation: Băroiu, A.-C.; Trăușan-Matu, Ș. Automatic Sarcasm Detection: Systematic Literature Review. *Information* **2022**, *13*, 399. https://doi.org/10.3390/info13080399

Academic Editor: Diego Reforgiato Recupero

Received: 27 June 2022
Accepted: 17 August 2022
Published: 22 August 2022

1. Introduction

Natural Language Processing (NLP) has been one of the most active fields of AI research in the past decade. Great strides have been made to bring machines closer to a human level understanding of language and, in many instances, the results have been groundbreaking. One area that has been very lucrative is sentiment analysis, the machine's ability to correctly identify the sentiment polarity of a statement or utteranc [1] Sentiment analysis is popular in both academia and industry, where it helps model trends and business strategies alike.

However, researchers have always encountered difficulties while performing sentiment analysis when figurative language forms are present, such as irony and sarcasm. These language instances are almost always used to convey the opposite meaning of what is said. The object of this paper, sarcasm, is defined by some as "a form of irony that is used in a hurtful or critical way [2] Others define it "as a subtype of verbal irony distinguished by the expression of a negative and critical attitude toward a recognizable victim or group of victims [3] Both definitions state that sarcasm requires the presence of a victim, towards which a negative sentiment, hurtful or critical, is addressed. Additionally, sarcasm is often described as a form of irony. Irony may be included in the domain of pragmatics, which relates to the role of context in conveying meaning.

Sarcasm can make data noisy. For example, "I love traffic!" is clearly a sarcastic sentence that expresses a negative sentiment. However, if a model is not fitted to account for sarcasm, it could deem that this sentence expresses a positive sentiment. To counter this noise, researchers developed models that can correctly identify the presence of sarcasm in target utterances. As such, automatic sarcasm detection has shaped a sub-area of sentiment analysis and NLP research. This paper will conduct a systematic literature review of this sub-area of research, automatic sarcasm detection, and will present its findings in the following sections. We consider that a systematic literature review will serve all researchers

in the Natural Language Processing field and beyond. They will be able to quickly assess the state-of-the-art data in the sarcasm detection research field and they will be able to more easily select approaches to tackle this task and others using the findings highlighted in this paper.

The paper is structured as follows: the next section will present the details about how the review was conducted and will present in summary the results of the review. The third section will present in depth the results of the review. The discussion section will offer a critique of the literature. The conclusions and future research section will postulate possible directions going forward and will summarize the paper and present final thoughts.

2. Materials and Methods

The research method used in this paper is a systematic literature review (SLR [4] SLR enables researchers to review current trends and future directions of study, as it allows multiple studies to be reviewed in a single grouping. To conduct this SLR, the PRISMA Guidelines were followed (https://prisma-statement.org/ (accessed on 10 August 2022)), which aids authors in improving the reporting of meta-analyses and systematic reviews.

The research questions of this review are:

1. What are the main areas of improvement that automatic sarcasm detection has seen since its inception?
2. What are the trends that have shaped the automated sarcasm detection task in the past decade?

Studying the existing literature on both sarcasm as a form of figurative language and sentiment analysis as a subarea of NLP, applying the knowledge in the research field and using the research questions as a starting point, the following search terms have been identified: automatic, sarcasm, irony, figurative language, detection, recognition, NLP, machine learning, deep learning, sentiment analysis and artificial intelligence (AI).

The following search strings, resulting from the search terms, were used:

("sarcasm" OR "irony" OR "figurative language") AND ("detection" OR "recognition")
("automatic") AND ("sarcasm" OR "irony" OR "figurative language") AND ("detection" OR "recognition")
("machine learning" OR "deep learning") AND ("sarcasm" OR "irony" OR "figurative language") AND ("detection" OR "recognition")
("NLP") AND ("sarcasm" OR "irony" OR "figurative language") AND ("detection" OR "recognition")
("AI") AND ("sarcasm" OR "irony" OR "figurative language") AND ("detection" OR "recognition")
("sentiment analysis") AND ("sarcasm" OR "irony" OR "figurative language")
("sentiment analysis") AND ("sarcasm" OR "irony" OR "figurative language") AND ("detection" OR "recognition").

The strings were used to search the following scientific databases: Web of Science, Science Direct, Google Scholar, Scopus and IEEE Xplore. These databases were selected because they are the largest on the internet and are guaranteed to have the noteworthy papers published in the subject of interest. Each database was searched separately between 20 May–5 July 2021.

There were several inclusion and exclusion criteria used in this study. First, to determine the year of publishing, the starting point was set by the paper largely credited to be the first to postulate and tackle the automatic sarcasm detection task. This paper was first published in 201 [5] There were prior attempts to detect sarcasm in text; however, they are often disregarded as being part of this area of research [6–8] As such, in this study there will only be papers published from 2010 onward. Another screening criterion was domain. This study included only the papers from the Computer Science domain. Therefore, papers that tackled sarcasm detection in an automatic, AI-related approach were considered. Papers from other domains, like Neuroscience, were not considered. The final criterion was the language that the papers were published in, with only documents available in

English being included. This criterion was set to avoid the complexity and confusion of translation. The papers that passed all these screening steps were then reviewed to verify their eligibility. First, the abstracts of the papers were analyzed to quickly identify if and how the papers tackled the automated sarcasm detection task. Then, the qualifying full papers were reviewed. This process was essential to ensure that only those papers that had sarcasm detection as their main scope were studied. This step eliminated those studies that used sarcasm detection alongside other topics. All the screening steps were carried out manually, and no automation tool was used during the conduction of this study.

As such, after searching the research databases presented prior with the selected search strings, 271 articles were identified. Of these, 142 were deemed irrelevant, with another 31 articles being excluded due to them not being in English or their type or title not being aligned with the subject matter. After this step, the abstracts of the remaining papers were studied and 27 were excluded because they did not align with the inclusion criteria of the study. Lastly, 11 papers were excluded following the full text review, as it was observed that sarcasm detection was not a central point of the research presented. After applying the steps described prior, 60 papers were identified. The inclusion and exclusion criteria by which the papers were selected are presented in Figure 1.

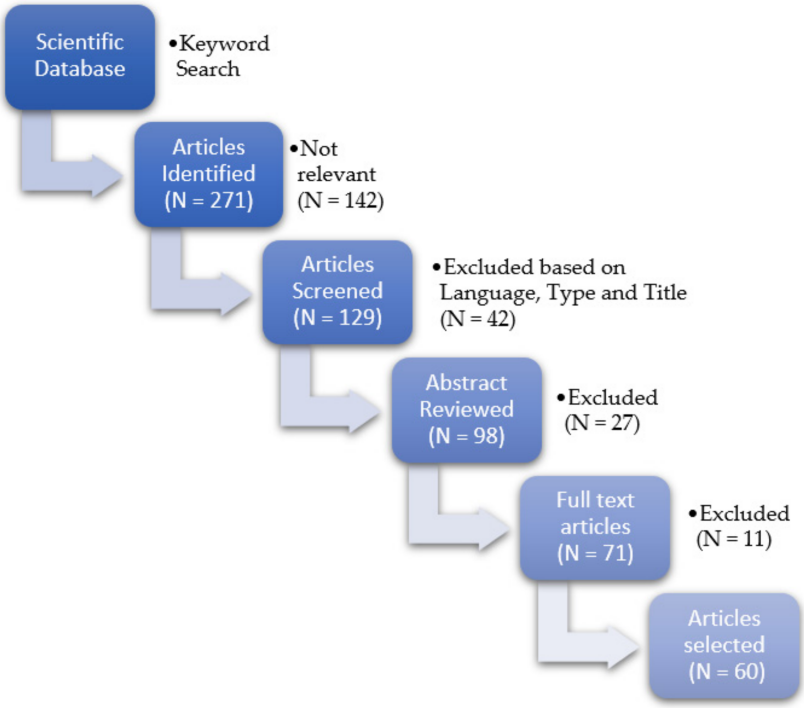

Figure 1. Inclusion and Exclusion Criteria, following the PRISMA Guidelines.

The earliest retrieved paper that is credited with tackling the automatic sarcasm detection task was published in 2010 and it was the only one published in that year. Interest for this area of research has steadily grown in the past decade, with more papers being published every year. This interest has led to the introduction of a new task in the SemEval competition, with the 2020 edition seeing an increased interest in Task 3, the sarcasm detection task. As such, a great number of papers on this topic were published in 2018 and 2020, years in which the SemEval took place. However, the positive trend of research articles in the field must be noted, even disregarding the SemEval competition. Ever since its inception, sarcasm detection has seen an increasing number of articles published every year, with a record 14 total articles identified for the year 2020. The trend of publication can be observed in Figure 2.

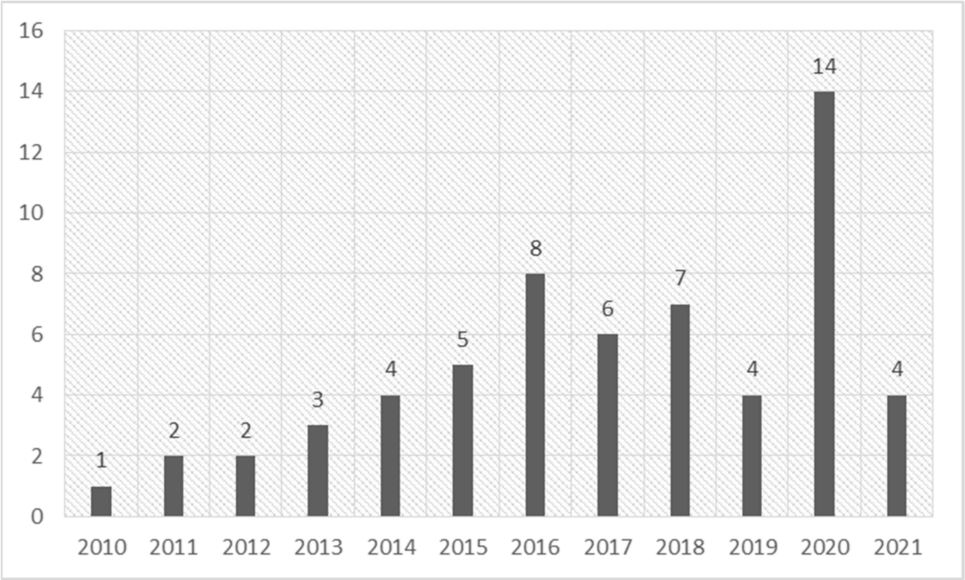

Figure 2. Publication trend by Year for Studies on Automatic Sarcasm Detection in English. Source: Own work. Note: 2021 until end of May.

3. Results

The thematic analysis of the papers highlights two main issues: sarcasm dataset creation and sarcasm detection. The first topic is mainly, but not exclusively, concerned with the introduction of a new dataset or a new rule by which a dataset can be constructed. The second topic is mainly, but not exclusively, concerned with the task of sarcasm detection and uses an established dataset or creates its dataset based on rules set by others. There are papers that tackle both issues.

3.1. Sarcasm Detection Datasets

First, the datasets and dataset generation rules used by researchers will be presented and analyzed. There are two approaches that are used when constructing a dataset: distant supervision and manual annotation.

Distant supervision is the quicker, more efficient way to build a dataset. It enables researchers to tap into the APIs of established social networks or websites, such as Twitter, Reddit or Amazon, and to collect millions of examples without any manual labor. Examples are considered positive or sarcastic when they meet a certain criterion, *#sarcasm* for Twitter or */s* for Reddit. Twitter is the most popular source of data and most dataset generation rules for it require the researcher to query the APIs for tweets that have one or more hashtags, such as *#sarcasm*, *#irony*, *#sarcastic*, *#not* or others. The data is then filtered by eliminating tweets where the hashtags are not at the end, retweets or non-English tweets [5,9–11].

Manual annotation uses crowdsourcing, employing human labor. A target utterance is presented to an annotator and they must state whether it is sarcastic or not. Until recently, the annotator was never the author of the utterance, therefore the labeling was based on perceived sarcas [12] A new dataset, iSarcasm, uses the authors of the utterances as the annotators, the labeling occurring based on intended sarcas [13]

A trend that is shaping for sarcasm datasets is the use of multiple data sources. While most papers prior to 2020 almost exclusively used Twitter datasets, more recent papers have displayed interest in multiple data sources, such as Reddit, news, books or even YouTube. A distribution of the dataset source for the identified papers is presented in Figure 3.

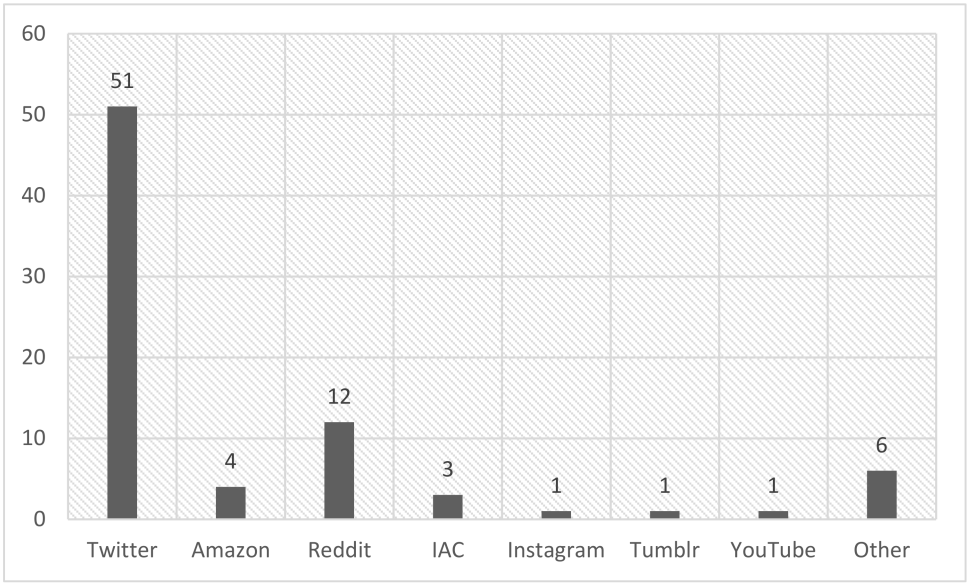

Figure 3. Dataset source distribution for the identified papers. Source: Own work.

It can be noted that Twitter is by far the most popular dataset source when it comes to automatic sarcasm detection, due to its popularity among English speakers, simple and concise text and the ease of extracting data through distant supervision by use of hashtags and APIs.

Recent trends for sarcasm datasets point toward special curated data and sets that are manually built or labeled, in favor of scrapping social media websites or forums. Researchers have opted for this approach to improve the quality of the data that is required to train better models. Sarcasm is known to be difficult to detect even under ideal conditions and noisy data that are usually found on the internet, for example Twitter, no longer suffice.

Next, the datasets used by the selected papers will be presented and analyzed. For each dataset used in each paper certain characteristics will be presented, inspired by the work of Joshi et al [14] There are three main metadata categories: annotation, context and dataset. The annotation category presents the method used to annotate the utterances as positive or negative. The two main techniques used are manual annotation and distant observation. For manual annotation a human judge has analyzed the utterance and labeled it accordingly. For distant observation the researchers used signals, such as Twitter hashtags, to label the data. For example, a tweet is considered sarcastic if it has #*sarcasm* in its text. The context is split into author and conversation context. The datasets that have author context contain information about the author of the target utterance, like past activity or profile information. Conversation context contains information about previous sentences in a conversation before the target utterance. The dataset data presents information about data type. Short datasets are composed of small texts, such as tweets, long datasets are composed of long texts, such as news articles or Reddit conversations, and other data represent non-text data, such as images or video data. The size represents the total number of instances present in the dataset. These characteristics for each dataset identified are presented in Table 1. The findings are further discussed in Section 4.

Table 1. Dataset information for the selected papers.

		Annotation		Context		Dataset			
Ref.	**Data Source**	**Manual**	**Distant**	**Author**	**Conversation**	**Short**	**Long**	**Other**	**Size**
[5]	Twitter		x			x			**5.9 million**
[5]	Amazon	x					x		**66,000**
[10]	Twitter		x			x			**900**
[11]	Amazon		x				x		**8861**
[15]	Amazon	x					x		**1254**
[16]	Twitter		x			x			
[17]	Internet Argument Corpus	x			x		x		**9889**
[12]	Twitter		x			x			**3000**
[18]	Twitter		x			x			**3.3 million**
[19]	Twitter	x				x			**134**
[20]	Twitter		x			x			**5.9 million**
[20]	Amazon		x				x		**8661**
[20]	News	x					x		**4233**
[21]	Twitter	x				x			**100,000**
[21]	Twitter		x			x			**4000**
[9]	Twitter		x			x			**60,000**
[22]	Twitter		x			x			
[23]	Twitter		x	x	x	x			**19,534**
[24]	Twitter		x			x			**5208**
[24]	Twitter		x			x			**2278**
[24]	IAC		x		x		x		**1502**
[25]	Twitter		x			x			**2.5 million**
[26]	Twitter		x			x			
[27]	Twitter		x	x		x			**Up to 1 million**
[28]	Twitter	x		x		x			**22,402 tweet conversations**
[29]	Twitter	x				x			**6000**
[30]	Twitter, Instagram and Tumblr	x				x		x	**44,010**
[31]	News		x				x		**4233**
[32]	Twitter		x		x	x			**9104**
[33]	Twitter		x			x			**50,000**
[34]	Twitter		x			x			**39,000**
[35]	Twitter		x			x			**650,000**
[36]	Twitter		x		x	x			**25,991**
[36]	IAC		x		x		x		**4692**
[37]	Twitter		x			x			**10,000**
[38]	Twitter		x			x			
[39]	Twitter		x			x		x	**843**
[40]	Twitter	x		x		x			**40,000**
[41]	Reddit	x		x	x		x		**Over 1 million**
[42]	Twitter		x			x			**45,000**
[42]	Reddit		x		x		x		**22,000**
[42]	IAC		x		x		x		**5000**
[43]	Amazon	x					x		**1254**
[44]	Twitter	x				x			**16,784**
[45]	Reddit	x			x		x		**Over 1.6 million**
[46]	Twitter	x				x			**3000**
[47]	Twitter	x		x	x	x			**8727**
[48]	Twitter	x		x		x			**701**
[48]	Twitter		x	x		x			**27,177**
[49]	YouTube	x		x	x			x	**690**

Table 1. *Cont.*

		Annotation		Context		Dataset			
Ref.	**Data Source**	**Manual**	**Distant**	**Author**	**Conversation**	**Short**	**Long**	**Other**	**Size**
[50]	Twitter		x			x			**51,189**
[51]	Twitter		x			x		x	**4819**
[52]	Twitter		x			x		x	**4819**
[53]	Twitter	x				x			**3000**
[53]	Twitter		x			x			**3000**
[53]	Reddit	x			x		x		**Over 1.6 million**
[54]	Twitter		x			x			**63,104**
[54]	IAC		x		x		x		**1935**
[54]	IAC		x		x		x		**4692**
[55]	Twitter		x			x			
[56]	Twitter		x			x			**6000**
[57]	Twitter		x		x	x			**5000**
[57]	Reddit		x		x		x		**4400 ***
[58]	Twitter		x		x	x			**5000 ***
[58]	Reddit		x		x		x		**4400 ***
[59]	Twitter		x		x	x			**5000 ***
[59]	Reddit		x		x		x		**4400 ***
[60]	Twitter		x		x	x			**5000 ***
[60]	Reddit		x		x		x		**4400 ***
[61]	Twitter		x		x	x			**5000 ***
[61]	Reddit		x		x		x		**4400 ***
[62]	Twitter		x		x	x			**5000 ***
[63]	Twitter		x		x	x			**5000 ***
[63]	Reddit		x		x		x		**4400 ***
[64]	Twitter		x		x	x			**5000***
[64]	Reddit		x		x		x		**4400 ***
[13]	Twitter	x		x		x		x	**4484**
[65]	Twitter		x			x			**1956**
[65]	Twitter		x			x			**54,931**
[65]	News		x				x		**26,709**
[66]	Twitter	x				x			**4618**
[67]	Twitter		x			x			**224**
[67]	Reddit		x				x		**950**
[67]	Books		x				x		**506**
[68]	Twitter	x				x			**3000**

(* same datasets used for the FigLang2020 workshop).

3.2. Automatic Sarcasm Detection

The approaches used to tackle automatic sarcasm detection have evolved throughout the years. They have transitioned from being rule- or feature-based to machine learning-based and, most recently, deep learning-based. An overview of the models used in the selected papers, in chronological order starting from 2010, is shown in Table 2.

Table 2. Overview of the models used in the selected papers. Source: Own work.

No.	Dataset Source	Model	Year	Ref.
1	Twitter and Amazon	kNN	2010 [5]	[5]
2	Twitter	SVM with sequential minimal optimization (SMO) and Logistic Regression	2011 [10]	[10]
3	Amazon	NB, SVM, DT	2011 [11]	[11]
4	Amazon	Dataset creation through manual annotation	2012 [15]	[15]
5	Twitter	Rule based	2012 [16]	[16]
6	IAC	Bootstrapped\High Precision and Pattern Based Classifiers	2013 [17]	[17]
7	Twitter	Bootstrapping	2013 [12]	[12]
8	Twitter	Balanced Window	2013 [18]	[18]
9	Twitter	GATE	2014 [19]	[19]
10	Amazon, Twitter and News	Maximum Entropy, Naïve Bayes and SVM	2014 [20]	[20]
11	Twitter	Maximum Entropy and SVM	2014 [21]	[21]
12	Twitter	Decision Tree	2014 [9]	[9]
13	Twitter	SCUBA: L1 regularized logistic regression	2015 [22]	[22]
14	Twitter	l2 regularization binary logistic regression	2015 [23]	[23]
15	Twitter	Feature-based statistical classifier	2015 [24]	[24]
16	Twitter	SVM with MVMEwe kernel (Maximum valued matrix element word embeddings)	2015 [25]	[25]
17	Twitter	Rule-based classifier	2015 [26]	[26]
18	Twitter	Rule-based classifier	2016 [27]	[27]
19	Twitter	Standard logistic regression with l2 regularization	2016 [28]	[28]
20	Twitter	Random forest, SVM, kNN and Maximum Entropy	2016 [29]	[29]
21	Twitter, Instagram and Tumblr	SVM and ANN	2016 [30]	[30]
22	News	Ensemble method	2016 [31]	[31]
23	Twitter	ANN	2016 [32]	[32]
24	Twitter	Naïve-Bayes, Decision Tree, Random Forest, SVM, Logistic Regression	2016 [33]	[33]
25	Twitter	SVM and CNN-DNN-LSTM	2016 [34]	[34]
26	Twitter and Forum	SVM and Naïve Bayes	2017 [35]	[35]
27	Twitter	SVM and LSTM	2017 [36]	[36]
28	Twitter	LSTM	2017 [37]	[37]
29	Twitter	SMO, BayesNet, J64	2017 [38]	[38]
30	Twitter	CNN	2017 [39]	[39]
31	Twitter	Word Embeddings—2 CNN—Bi-LSTM—DNN	2017 [40]	[40]
32	Reddit	CASCADE (Context and Content Features, CNN)	2018 [41]	[41]
33	Twitter	MIARN	2018 [42]	[42]
34	Twitter	SVM and Random Forest	2018 [43]	[43]
35	Twitter	Naïve-Bayes, SVM, Decistion Tree, Random Forest, Adabosst, kNN, Gradient Boost	2018 [44]	[44]
36	Reddit	Dataset creation through distant supervision	2018 [45]	[45]
37	Twitter	Multiple models (SemEval 2018 submissions)	2018 [46]	[46]
38	Twitter	CANN-KEY and CANN-ALL	2018 [47]	[47]
39	Twitter	Exclusive and inclusive models with Cascade, Encoder decoder and summary embeddings	2019 [48]	[48]
40	YouTube	SVM (dataset creation-oriented)	2019	[49]
41	Twitter	Naïve-Bayes	2019 [50]	[50]
42	Twitter	Multimodal ANN	2019 [51]	[51]
43	Twitter	Multimodal ANN	2020 [52]	[52]
44	Twitter	Recurrent CNN RoBERTA	2020 [53]	[53]
45	Twitter	MMNSS	2020 [54]	[54]
46	Twitter	Feature-based statistical model	2020 [55]	[55]
47	Twitter	BiLStM—CNN	2020 [56]	[56]
48	Twitter and Reddit	BiLSTM, BERT and SVM	2020 [57]	[57]
49	Twitter and Reddit	BERT, RoBERTA, spanBERT	2020 [58]	[58]
50	Twitter and Reddit	BERT 3,5,7, all	2020 [59]	[59]
51	Twitter and Reddit	Ensemble method	2020 [60]	[60]

Table 2. *Cont.*

No.	Dataset Source	Model	Year	Ref.
52	Twitter and Reddit	Bert-large+BiLSTM+NextVLAD	2020 [61]	[61]
53	Twitter and Reddit	Ensemble method	2020 [62]	[62]
54	Twitter and Reddit	BERT-CNN-LSTM	2020	[63]
55	Twitter and Reddit	RoBERTA	2020 [64]	[64]
56	Twitter	Dataset creation through manual annotation	2020 [13]	[13]
57	Twitter and News	HA-LSTM	2021 [65]	[65]
58	Twitter	CNN	2021 [66]	[66]
59	Twitter, Reddit and Books	Ensemble method	2021 [67]	[67]
60	Twitter	BERT	2021 [68]	[68]

From Table 2 it can be noted that automatic sarcasm detection has followed the trends that have shaped NLP research in the past decade. The first papers published in this area focused on machine learning and feature-based models. Classifiers such as Support Vector Machine [69], Logistic Regression, Decision Tre [70] Naïve Bayes and Random Fores [71] dominated the landscape in the first years of the decade. Then, a shift began toward deep neural networks, such as Convolutional Neural Network [72] Long Short-Term Memor [73] and other different configurations shaping the progress for sarcasm detection. Recent years have seen the surge of transformers, with BER [74] and RoBERTA [75] setting the path for new advancements in the field. Additionally, the analysis of these papers has highlighted new trends in the field, multi-modal approaches becoming more popular with the integration of deep learning.

Next, the reported performance of the selected articles will be presented and analyzed. Because sarcasm detection is a classification problem, the metrics used by researchers are Precision, Accuracy, Recall, F1-score and Area Under Curve (AUC). The performance information is found in Table 3. The findings are further discussed in Section 4.

Table 3. Reported performance for each identified paper.

Reference	Precision	Accuracy	Recall	F-Score	AUC	Data
[5]	72.7	43.6	89.6	54.5		Twitter
[5]	91.2	75.6	94.7	82.7		Amazon
[10]		71				Twitter
[11]	77.1	75.75	72.5	74.7		Amazon
[16]	78	70.1	56	65		Twitter
[17]	62		52	57		IAC
[12]	62		44	51		Twitter
[18]					79	Twitter
[19]	77.3		77.3	77.3		Twitter
[20]					84	Twitter
[20]					85.4	Amazon
[20]					83.3	News
[21]				94.66		Twitter
[21]				58.2		Twitter
[9]	62		90	90		Twitter
[22]		86.1			86	Twitter
[23]		85.1				Twitter
[24]	81.4		97.6	88.8		Twitter
[24]	77		51	61		Twitter
[24]	48.9		92.4	64		IAC
[25]	96.6		98.5	97.5		Twitter
[26]	85		96	90		Twitter
[27]	97		98	97		Twitter
[28]					63	Twitter
[29]	91.1	83.1	73.4	81.3		Twitter

Table 3. *Cont.*

Reference	Precision	Accuracy	Recall	F-Score	AUC	Data
[30]		69.7				Twitter
[30]		74.2				Instagram
[30]		70.9				Tumblr
[31]	94.4			95.5	97.4	News
[32]		90.74		90.74		Twitter
[33]				58.6		Twitter
[34]	91.9		92.3	91.2		Twitter
[35]	96.5	65.2	20.10	37.4		Twitter
[36]	77.25		75.51	76.36		Twitter
[36]	66.9		82.1	73.7		IAC
[37]	85.5	85.5	85.5	85.5		Twitter
[38]				75		Twitter
[39]	87.1		86.9	86.97		Twitter
[40]	90		89	90		Twitter
[41]		79		86		Reddit
[42]	86.1	86.5	85.8	86		Twitter
[42]	69.7	69.9	69.4	69.5		Reddit
[42]	72.9	72.8	72.9	72.8		IAC
[43]	80.1		73.6	75.2		Amazon
[44]		93		92		Twitter
[46]	63	73.5	80.1	70.5		Twitter
[47]				63.3		Twitter
[48]				73.9		Twitter
[48]				93.4		Twitter
[49]	72.1		71.7	71.8		YouTube
[50]				89.8		Twitter
[51]	76.6	83.4	84.2	80.2		Twitter
[52]	80.9	86.1	85.1	82.9		Twitter
[53]	81	82	80	80	89	Twitter
[53]	90	91	90	90	94	Twitter
[53]	78	79	78	78	85	Reddit
[54]			89.2	87.1		Twitter
[54]	75		70.9	67.7		IAC
[54]	85.8		71.1	74.2		IAC
[55]	93.8	93.5				Twitter
[56]	78.8	85	81.3	79.5		Twitter
[57]	74.4		74.8	74.3		Twitter
[57]	65.8		65.8	65.8		Reddit
[58]	77.3		77.4	77.2		Twitter
[58]	69.3		69.9	69.1		Reddit
[59]				75.2		Twitter
[59]				62.1		Reddit
[60]	74.1		74.6	74		Twitter
[60]	67		67.7	66.7		Reddit
[61]	93.2		93.6	93.1		Twitter
[61]	83.4		83.8	83.4		Reddit
[62]	79		79.2	79		Twitter
[63]				74		Twitter
[63]				63.9		Reddit
[64]	77.2		77.2	77.2		Twitter
[64]	71.6		71.8	71.6		Reddit
[65]				99		Twitter
[65]				98		Twitter
[65]				88		News
[66]				66		Twitter
[67]		54.9		21.7		Twitter
[67]		60.3		46.2		Reddit
[67]		50.8		18.5		Books
[68]	68.7	70.6	72.5	70.5		Twitter

4. Discussion

As can be seen from Table 2, the early years of sarcasm detection were dominated by machine learning models. The first robust algorithm used for sarcasm detection, the semi-supervised sarcasm identification algorithm (SASI) to detect sarcasm in Twitter and Amazon product reviews [5], appeared in 2010. At the time, all systems failed to correctly classify the sentiment of sarcastic sentences. This algorithm used two modules: semi-supervised pattern acquisition to identify sarcastic patterns that could be used as features in a classifier and a classification stage to assign each sentence to a sarcastic class. The authors of [10] studied lexical and pragmatic features in tweets, using unigrams and dictionary-based for classifying sarcastic, positive and negative tweets by employing two classifiers: SVM and logistic regression.

From the analyzed papers, there are interesting approaches that must be noted. One such approach is the SCUBA framewor [22] The authors wanted to improve sarcasm detection on Twitter by integrating past tweets of the target tweet's author. The authors developed several features that derived from forms that sarcasm can take, such as contrast of sentiments, the form of written expression, the means of conveying emotion and others. They trained several models and chose L1-regularized logistic regression as the preferred option. The results showed that the accuracy of the predictions increased as the number of past tweets that the model has access to also increased. Other researchers also accounted for contex [23] They extracted several features that could give information about said context and split these features in three categories: tweet, author and audience features. They used binary-logistic regression with L2 regularization to classify the texts. The results showed that the author features were the salient features, the performance of the classifier improving almost as much as when all the features were introduced in the model. Additionally, the authors found that *#sarcasm* was used by the tweet authors when they were not familiar with their audience and wanted to make sure their message was correctly perceived. These papers highlight that when context is accounted for, the performance of sarcasm detection models increases.

Some researchers introduced emojis into the mix [37]. The authors employed a deep learning approach and trained their own word embeddings to properly capture the salient information provided by emojis. The DeepMoji model that the authors proposed consists of an embedding layer that feeds in two BiLSTM layers that feed into an attention layer and a final softmax activation function that makes the prediction. The results showed that the diversity of the emojis used is important to the performance of the model.

Other interesting approaches integrated English with other languages, like Cantonese [20] or HindI [56]. The Chinese model first extracted sarcasm features from Cantonese and English texts, then applied weighted random sampling to these texts, followed by bagging. A weighted vote was applied to find the best classifier. The Indian model consisted of three parts, one that uses an attention-based BiLSTM that generates context vectors for English, one that uses Hindi-SentiWordNet to generate feature vectors for Hindi and a classifier, which is trained on three features, English, Hindi and auxiliary pragmatic features. Sarcasm detection has also been implemented to counter cyberbullyin [66] The study found similarities between ironic and sarcastic tweets. Even more, sarcasm was found to be a great indicator for the presence of cyberbullying, further proving the practical applicability of sarcasm in NLP tasks.

Oprea and Mand [48] defined author context as the embedded representation of their historical Twitter posts and proposed neural models for extracting these representations. They tested two tweet datasets, one manually labeled for sarcasm and the other using tag-based distant supervision. Exclusive models in the authors' proposed architecture did not use the current tweet being classified, instead basing the prediction solely on user history. In contrast, inclusive models took into account both user history and the most recent tweet.

Multimodal approaches must also be noted, due to their increased popularity in recent years. The first multimodal approach integrated images in the sarcasm detection

task [30]. The authors collected data from three social media platforms, Twitter, Instagram and Tumblr. They then employed two approaches to sarcasm detection, a SVM approach and a deep learning approach. For the SVM approach, they extracted NLP and visual semantics features. For the DL approach, they used two networks, an NLP one and a visual one, which they then fused in order to achieve the prediction. The results showed that integrating visual information improved performance for the Instagram set, while it was inconsequential for Twitter and Tumblr. Additionally, text features proved to offer little for the performance of the deep learning approach. Another approach [51] proposed a hierarchical fusion model that implemented three feature representations: image, image attribute and text. The authors of the paper treated text features, image features and image attributes as three modalities. The proposed hierarchical fusion model in the paper extracted image features and attribute features first and then used attribute features and a bidirectional LSTM (BiLSTM) network to extract text features. After that, the model reconstructed the features of three modalities and fused them into a single feature vector for prediction. The authors trained and tested their approach on a multi-modal dataset based on Twitter.

A multimodal approach [52] based on BERT for text preprocessing was also proposed. The study was conducted on Twitter data that had both text and image. The model integrated three components, text, hashtag and image. The model made use of both inter-modality attention, between image and text, and intra-modality attention, within the text.

One interesting observation is that the best performing solutions in the SemEval 2020 used ensemble methods and/or implemented data augmentation. Therefore, semi-supervised techniques in conjunction with transformer-based architectures could attain superior results over other approaches and should be favored going forward.

There are multiple observations to be made from Table 1. First, the variability in dataset size must be noted. With the exception of the FigLang2020 workshop, most papers use different sized datasets. Even the papers that try to use established sarcasm datasets, similarly to the Riloff tweet dataset, encounter difficulties. Because these datasets were constructed through Twitter API and only tweet ids were given, the longer the time passes, the more the datasets deteriorate. Due to Twitter change of policies, the users deleting the tweets or other events, the datasets become smaller and information is lost. This size variability makes performance benchmarking more difficult because parity is lost. For some papers the dataset size information is missing altogether. The better reporting of the dataset used and its characteristics could be employed by future researchers.

Next, it can be seen that most datasets are annotated through distant observation. Forty-six of the unique datasets identified are constructed this way, more than double those of manual annotation, i.e., twenty. Research has shown that manual annotation is superior to distant observation and future research should focus more on building and working with manually annotated datasets. Context is also found to be lacking, especially in older datasets. Only 10 unique datasets have author context and 18 have conversation context. Again, future research should focus on building and working with datasets that include context information, if better performing solutions are to be developed. On the topic of data type, most datasets are composed of short texts (44), fewer are composed of long texts (19) and only 6 texts include non-text data. Multi-modal approaches have generated increased interest in recent years and more datasets that incorporate different types of data, like MUStARD, should be developed.

There is valuable information that can be extracted from Table 3. At first glance, the performance boost of deep learning can be observed. Solutions proposed from 2016 onward see an increase in metrics scores. However, the scores do not tell the whole story. As seen in Table 2, the great variability of datasets or size of the same dataset make performance benchmarking a difficult task. Past solutions were able to be trained on complete datasets and achieve equal or superior performance to more recent implementations, especially if the proposed approach is data hungry, such as transformer-based architectures. The performance of modern solutions is, however, superior to past approaches. This difference

is highlighted in the FigLang 2020 workshop where the winning proposal netted excellent results, far superior to any past solution.

However, this study has identified some key issues with the automatic sarcasm detection task. These issues spring from the datasets and dataset creation rules. First, for distant supervision, the dataset ends up being noisy. The assumption is made that sarcasm is present only in instances that have a certain identifier, such as *#sarcasm* or */s*. This is untrue and sarcastic instances end up in the false class, only because they lacked the identifier, and therefore lead to false negatives. Furthermore, this process only captures one type of sarcasm, one that is specific to a clearly established setting, a Twitter or a Reddit thread. This limits the ability of models to identify other flavors of sarcasm and therefore leads to false negatives.

Manual annotation methods have also proven to create less than ideal datasets. One crucial problem is perceived vs. intended sarcasm. For almost all datasets built this way, the annotator is different from the author of the target utterance. This can lead to a low agreement rating between the author and the annotator and can lead to both false positives and false negatives. Training a model on these sets is akin to training it on perceived sarcasm. To counter this, iSarcasm has the authors of a target utterance to annotate it. Therefore, the utterance is correctly labeled by its author.

However, the problem of perceived vs. intended sarcasm does not go away. Training a model on this dataset will simply shift the perception to intended sarcasm. Research has shown that different cultures perceive sarcasm differently or second language speakers have difficulties understanding sarcasm [76,77]. As such, a dataset that is skewed toward each part can prove detrimental to universal sarcasm detection, if such a task can even be performed.

Furthermore, sarcasm detection has traditionally proven to be a very difficult task, even for humans. By relying heavily on a single source, Twitter, the different flavors and facets of sarcasm are lost. Recent approaches, such as the one by Castro et al. [49], must be encouraged. The dataset that the authors propose is collected from sitcoms and has text, audio and video data. However, it is not without fault. The nature of sitcoms is to exaggerate situations and purposefully make jokes, therefore the sarcasm present in them tends to be different from the one used daily by humans. As such, better sources of data for sarcasm must be found or created.

As previously stated, the models used in the automatic sarcasm detection task have evolved throughout the years. NLP trends have shaped the task, with the large adoption of deep learning starting in 2016 and transformers shaping the landscape in 2020 and 2021. For example, all papers published from 2020 onward have included transformers. Their impact cannot be understated, and the automatic sarcasm detection task has benefited greatly from their introduction. However, it can be said that the problem has been wrongly defined. If the goal is to aid corporations to identify sarcastic tweets and correctly respond, then the path might be right. However, if the goal of NLP research is to build models that can replicate human level ability, then sarcasm detection still has a long way to go.

Recent trends in the automatic detection of sarcasm point towards multi-modal, deep learning approaches. In recent years, researchers have shifted their focus to implementing transformer-based architectures in order to tackle this task. As seen in other areas of NLP, transformers have brought a new era of performance and have quickly achieved the state-of-the-art. However, because multiple, heterogeneous datasets are used for the sarcasm detection task, benchmarking these models remains difficult. Going forward, a handful of datasets could be identified that would serve for benchmarking purposes and that would help the development of the research field. Both MUStARD and iSarcasm could be great starting points for selecting such datasets.

The systematic literature review allowed us to cover a broad field of research and to extract valuable information that was presented and discussed in the previous sections. However, this study has some limitations. Future researchers could use the same method-

ology and query more databases. Additionally, they could introduce more search terms in their query.

This study has also disregarded any papers that were not published in English and, therefore, ignored important research on sarcasm detection in other languages. Due to its nature, being time-bound, this study can always be replicated in the future to assess the progress that has been made in the field.

5. Conclusions and Future Research

Stemming from the findings presented in the previous section, a few directions for future research will be presented. First and, we believe, most importantly, researchers should investigate better ways to construct their datasets. Twitter can be a good source of data, but it must not be the only one. Researchers fight an uphill battle in this regard, with few networks providing access to their data, but a more varied approach must be considered. Multimodal datasets can also be further studied, as they tend to capture more of the nuances of sarcasm. Therefore, the following questions are asked:

a. What does automatic sarcasm detection want to achieve?
b. What are the best data to train models on to replicate human level ability?

Second, as sarcasm has proven to be heavily dependent on context, sarcasm context detection could be explored. Researchers could develop models that could correctly identify whether a certain context is appropriate for sarcasm to be used and, therefore, develop speech systems that could correctly use it. This could lead to more natural human–machine communication, as sarcasm is an integral part of human culture. This leads to the following question:

c. What is the correct context in which to use sarcasm?

Third, future researchers could explore the relation between different languages and cultures regarding sarcasm. They could develop models that can correctly translate or interpret a sarcastic remark from one language to another and identify if a certain context is appropriate for sarcasm in multiple languages. Interesting systems that can improve our understanding of sarcasm can also be developed. Such systems could transform normal utterances into sarcastic ones or vice versa, akin to a sarcasm translator that could decode the intent of someone like Chandler, a highly sarcastic character, from the "Friends" TV show. Such a system could lead to a better understanding of sarcasm and its application in a machine environment. This leads to the following question:

d. How can machines generate sarcasm?

After analyzing the selected papers, the two research questions considered in this paper can be answered as follows: the main area that automatic sarcasm detection has seen improvement in is the models. They have evolved in tandem with all of NLP research, from machine learning to transformers. Additionally, recent trends in the field have been identified, for both datasets and methods. However, one area that has seen slow improvement is the selection of data for the task. It has remained mostly unchanged for the past decade, and this has proven to be a problem. A re-evaluation of the task could be carried out by researchers and future avenues could be explored, especially regarding the data selection process.

However, this should not discourage future researchers. Sarcasm is a beautiful characteristic of human language and culture, and its application to a machine environment was never going to be easy. This review serves only as an assessment of the work carried out and peeks into a future where humans and machines can get along even better than today.

Author Contributions: Conceptualization, A.-C.B. and Ș.T.-M.; methodology, A.-C.B.; software, A.-C.B.; validation A.-C.B. and Ș.T.-M.; formal analysis, A.-C.B.; investigation, A.-C.B.; resources, A.-C.B. and Ș.T.-M.; data curation, A.-C.B.; writing—original draft preparation, A.-C.B.; writing—review and editing, A.-C.B. and Ș.T.-M.; visualization, A.-C.B.; supervision, Ș.T.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Authors can confirm that all relevant data are included in the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dang, C.; Moreno-Garcia, M.; De la Prieta, F. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics* **2020**, *9*, 483. [CrossRef]
2. McDonald, S.; Pearce, S. Clinical insights into pragmatic theory: Frontal lobe deficits and sarcasm. *Brain Lang.* **1996**, *53*, 81–104. [CrossRef] [PubMed]
3. Lee, C.; Katz, A. The differential role of ridicule in sarcasm and irony. *Metaphor Symb.* **1998**, *13*, 1–15. [CrossRef]
4. Raharjana, I.; Siahaan, D.; Fatichah, C. User Stories and Natural Language Processing: A Systematic Literature Review. *IEEE Access* **2021**, *9*, 53811–53826. [CrossRef]
5. Davidov, D.; Tsur, O.; Rappoport, A. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Uppsala, Sweden, 15–16 July 2010.
6. Carvalho, P.; Sarmento, L.; Silva, M.; de Oliveira, E. Clues for detecting irony in user-generated contents: Oh ... !! it's so easy;-). In Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, Hong Kong, China, 6 November 2009.
7. Tepperman, J.; Traum, D.; Narayanan, S. Yeah right: Sarcasm recognition for spoken dialogue systems. In Proceedings of the InterSpeech ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
8. Burfoot, C.; Baldwin, T. Automatic satire detection: Are you having a laugh? In *ACLShort '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing Short Papers, Suntec, Singapore, 2–7 August 2009*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009.
9. Barbieri, F.; Saggion, H.; Ronzano, F. Modelling sarcasm in twitter, a novel approach. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Baltimore, MD, USA, 27 June 2014.
10. González-Ibánez, R.; Muresan, S.; Wacholder, N. Identifying sarcasm in Twitter: A closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; Volume 2.
11. Reyes, A.; Rosso, P. Mining subjective knowledge from customer reviews: A specific case of irony detection. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), Portland, OR, USA, 24 June 2011.
12. Riloff, E.; Surve, P.; Qadir, A.; De Silva, L.; Gilbert, N.; Huang, R. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; Volume 13, pp. 704–714.
13. Oprea, S.; Magdy, W. iSarcasm: A Dataset of Intended Sarcasm. *arXiv* **2019**, arXiv:1911.03123v2.
14. Joshi, A.; Bhattacharyya, P.; Carman, M. Automatic Sarcasm Detection: A Survey. *ACM Comput. Surv.* **2018**, *50*, 1–22. [CrossRef]
15. Filatova, E. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In Proceedings of the 12th Language Resources and Evaluation Conference, Istanbul, Turkey, 23–25 May 2012.
16. Reyes, A.; Rosso, P.; Buscaldi, D. From humor recognition to irony detection: The figurative language. *Data Knowl. Eng.* **2012**, *74*, 1–12. [CrossRef]
17. Lukin, S.; Walker, M. Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue. In Proceedings of the Workshop on Language Analysis in Social Media, Atlanta, GA, USA, 13 June 2013.
18. Liebrecht, C.; Kunneman, F.; van den Bosch, A. The perfect solution for detecting sarcasm in tweets# not. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), Atlanta, GA, USA, 14 June 2013; Volume 2013.
19. Maynard, D.; Greenwood, M. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014.
20. Liu, P.; Chen, W.; Ou, G.; Wang, T.; Yang, D.; Lei, K. Sarcasm Detection in Social Media Based on Imbalanced Classification. In Proceedings of the International Conference on Web-Age Information Management, Macau, China, 15–16 June 2014.
21. Ptacek, T.; Habernal, I.; Hong, J. Sarcasm Detection on Czech and English Twitter. In Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland, 23–29 August 2014; pp. 213–223.
22. Rajadesingan, A.; Zafarani, R.; Liu, H. Sarcasm detection on Twitter: A behavioral modeling approach. In Proceedings of the WSDM 2015—Proceedings of the 8th ACM International Conference on Web Search and Data Mining, Shanghai, China, 31 January 2015.

23. Bamman, D.; Smith, N. Contextualized sarcasm detection on twitter. In Proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015.
24. Joshi, A.; Sharma, V.; Bhattacharyya, P. Harnessing context incongruity for sarcasm detection. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Beijing, China, 26–31 July 2015; pp. 757–762.
25. Ghosh, D.; Guo, W.; Muresan, S. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015.
26. Bharti, S.; Babu, K.; Jena, S. Parsing-based sarcasm sentiment recognition in Twitter data. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–28 August 2015.
27. Bharti, S.; Vachha, B.; Pradhan, R.; Babu, K.; Jena, S. Sarcastic sentiment detection in tweets streamed in real time: A big data approach. *Digit. Commun. Netw.* **2016**, *2*, 108–121. [CrossRef]
28. Abercrombie, G.; Hovy, D. Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twitter Conversations. In Proceedings of the ACL 2016 Student Research Workshop, Berlin, Germany, 7–12 August 2016.
29. Bouazizi, M.; Otsuki, T. A Pattern-Based Approach for Sarcasm Detection on Twitter. *IEEE Access* **2016**, *4*, 5477–5488. [CrossRef]
30. Schifanella, R.; de Juan, P.; Tetreault, J.; Cao, L. Detecting Sarcasm in Multimodal Social Platforms. In Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016.
31. Ravi, K.; Vadlamani, R. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowl. -Based Syst.* **2016**, *120*, 15–33. [CrossRef]
32. Zhang, M.; Zhang, Y.; Fu, G. Tweet Sarcasm Detection Using Deep Neural Network. In Proceedings of the COLING 2016, 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016.
33. Sulis, E.; Farias, D.; Rosso, P.; Patti, V.; Ruffo, G. Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm and #not. *Knowl.-Based Syst.* **2016**, *108*, 132–143.
34. Ghosh, A.; Veale, T. Fracking Sarcasm Using Neural Network. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, San Diego, CA, USA, 16 June 2016.
35. Saha, S.; Yadav, J.; Ranjan, P. Proposed Approach for Sarcasm Detection in Twitter. *Indian J. Sci. Technol.* **2017**, *10*, 1–8. [CrossRef]
36. Ghosh, D.; Fabbri, A.; Muresan, S. The Role of Conversation Context for Sarcasm Detection in Online Interactions. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2017), Saarbrücken, Germany, 15–17 August 2017.
37. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017.
38. Salas-Zarate, M.d.P.; Paredes-Valverde, M.; Rodriguez-Garcia, M.; Valencia-Garcia, R.; Alor-Hernandez, G. Automatic Detection of Satire in Twitter: A psycholinguistic-based approach. *Knowl.-Based Syst.* **2017**, *128*, 20–33. [CrossRef]
39. Mishra, A.; Dey, K.; Bhattacharyya, P. Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification using Convolutional Neural Network. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017.
40. Ghosh, A.; Veale, T. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017.
41. Hazarika, D.; Hazarika, S.; Gorantla, S.; Cambria, E.; Zimmermann, R.; Mihalcea, R. CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018.
42. Tay, Y.; Luu, A.; Hui, S.; Su, J. Reasoning with sarcasm by reading in in-between. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018.
43. Kumar, H.M.K.; Harish, B. Sarcasm classification: A novel approach by using Content Based Feature Selection Method. *Procedia Comput. Sci.* **2018**, *143*, 378–386. [CrossRef]
44. Ahuja, R.; Bansal, S.; Prakash, S.; Venkataraman, K.; Banga, A. Comparative Study of Different Sarcasm Detection Algorithms Based on Behavioral Approach. *Procedia Comput. Sci.* **2018**, *143*, 411–418. [CrossRef]
45. Khodak, M.; Saunshi, N.; Vodrahalli, K. A Large Self-Annotated Corpus for Sarcasm. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, 7–12 May 2018.
46. Van Hee, C.; Lefever, E.; Hoste, V. SemEval-2018 task 3: Irony detection in English tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), New Orleans, LA, USA, 5–6 June 2018; pp. 39–50.
47. Ren, Y.; Ji, D.; Ren, H. Context-augmented convolutional neural networks for twitter sarcasm detection. *Neurocomputing* **2018**, *308*, 1–7. [CrossRef]
48. Oprea, S.; Magdy, W. Exploring Author Context for Detecting Intended vs Perceived Sarcasm. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
49. Castro, S.; Hazarika, D.; Perez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; Poria, S. Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper). In Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
50. Recupero, D.; Alam, M.; Buscaldi, D.; Grezka, A.; Tavazoee, F. Frame-Based Detection of Figurative Language in Tweets. *IEEE Comput. Intell.* **2019**, *14*, 77–88. [CrossRef]

51. Cai, Y.; Cai, H.; Wan, X. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
52. Pan, H.; Lin, Z.; Qi, Y.; Fu, P.; Wang, W. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; Volume 2020.
53. Potamias, R.; Siolas, G.; Stafylopatis, A. A transformer-based approach to irony and sarcasm detection. *Neural Comput. Appl.* **2020**, *32*, 17309–17320. [CrossRef]
54. Ren, L.; Xu, B.; Lin, H.; Liu, X.; Yang, L. Sarcasm Detection with Sentiment Semantics Enhanced Multi-level Memory Network. *Neurocomputing* **2020**, *401*, 320–326. [CrossRef]
55. Sonawane, S.; Kolhe, S. TCSD: Term Co-occurrence Based Sarcasm Detection from Twitter Trends. *Procedia Comput. Sci.* **2020**, *167*, 830–839. [CrossRef]
56. Jain, D.; Kumar, A.; Garg, G. Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. *Appl. Soft Comput. J.* **2020**, *91*, 106198. [CrossRef]
57. Baruah, A.; Das, K.; Barbhuiya, F.; Dey, K. Context-aware sarcasm detection using BERT. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
58. Kumar, A.; Anand, V. Transformers on sarcasm detection with context. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
59. Avvaru, A.; Vobilisetty, S.; Mamidi, R. Detecting sarcasm in conversation context using Transformer based model. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
60. Lemmens, J.; Burtenshaw, B.; Lotfi, E.; Markov, I.; Daelemans, W. Sarcasm detection using an ensemble approach. In Proceedings of the Second Workshop on Figurative Language Processsing, Seattle, WA, USA, 9 July 2020.
61. Lee, H.; Yu, Y.; Kim, G. Augmenting data for sarcasm detection with unlabeled conversation context. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
62. Jaiswal, N. Neural sarcasm detection using conversation context. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
63. Srivastava, H.; Varshney, V.; Kumari, S.; Srivastava, S. A novel hierarchical BERT architecture for sarcasm detection. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
64. Dadu, T.; Pant, K. Sarcasm detection using context separators in online discourse. In Proceedings of the Second Workshop on Figurative Language Processing, Seattle, WA, USA, 9 July 2020.
65. Pandey, R.; Kumar, A.; Singh, J.; Tripathi, S. Hybrid attention-based Long Short-Term Memory network for sarcasm identification. *Appl. Soft Comput. J.* **2021**, *106*, 107348. [CrossRef]
66. Chia, Z.; Ptaszynski, M.; Masui, F.; Leliwa, G.; Wroczynski, M. Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Inf. Process. Manag.* **2021**, *58*, 102600. [CrossRef]
67. Parameswaran, P.; Trotman, A.; Liesaputra, V.; Eyers, D. Detecting the target of sarcasm is hard: Really? *Inf. Process. Manag.* **2021**, *58*, 102599. [CrossRef]
68. Shrivastava, M.; Kumar, S. A pragmatic and intelligent model for sarcasm detection in social media text. *Technol. Soc.* **2021**, *64*, 101489. [CrossRef]
69. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
70. Quinlan, J. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
71. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
72. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
73. Hochreitter, S.; Schimdhuber, J. Long-short term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
74. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019.
75. Liu, Y.; Zhou, Y.; Dong, F.; Wang, C.; Wang, Z. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering* **2019**, *5*, 156–163. [CrossRef]
76. Kim, J. How Korean EFL learners understand sarcasm in L2 English. *J. Pragmat.* **2014**, *60*, 193–206. [CrossRef]
77. Ackerman, B.P. Contextual integration and utterance interpretation: The ability of children and adults to interpret sarcastic utterances. *Child Dev.* **1982**, *53*, 1075–1083. [CrossRef]