Artificial Intelligence Review (2020) 53:4215–4258
https://doi.org/10.1007/s10462-019-09791-8

Sarcasm identification in textual data: systematic review, research challenges and open directions

Christopher Ifeanyi Eke[1,2] · Azah Anir Norman[1] · Liyana Shuib[1] · Henry Friday Nweke[1,3]

Published online: 30 November 2019

Abstract
Sarcasm is a form of sentiment whereby people express the implicit information, usually the opposite of the message content in order to hurt someone emotionally or criticise something in a humorous way. Sarcasm identification in textual data, being one of the hardest challenges in natural language processing (NLP), has recently become an interesting research area due to its importance in improving the sentiment analysis of social media data. A few studies have carried out a comprehensive literature review on sarcasm identification in the existing primary study within the last 11 years. Thus, this study carried out a review on the classification techniques for sarcasm identification under the aspects of datasets, pre-processing, feature engineering, classification algorithms, and performance metrics. The study has considered the published article from the period of 2008 to 2019. Forty (40) academic literature were selected from the 7 standard academic databases in order to carry out the review and realize the objectives. The study revealed that most researchers created their own datasets since there is no standard available datasets in the domain of sarcasm identification. Context and content-based linguistic features were used in most of the studies. This review shows that n-gram and parts of speech tagging techniques were the most commonly used feature extraction techniques. However, binary representation and term frequency were utilized for feature representation whereas Chi squared and information gain were used for the feature selection scheme. Moreover, classification algorithm such as support vector machine, Naïve Bayes, random forest, maximum entropy, and decision tree algorithm were mostly applied using accuracy, precision, recall and F-measure for performance measures. Finally, research challenges and future direction are summarized in this review. This review reveals the impact of sarcasm identification in building effective product reviews and would serve as handle resources for researchers and practitioners in sarcasm identification and text classification in general.

Keywords Sarcasm identification · Social media data · Natural language processing · Pre-processing · Feature engineering · Textual classification · Performance measure

✉ Christopher Ifeanyi Eke
ciekecsc@siswa.um.edu.my

✉ Azah Anir Norman
azahnorman@um.edu.my

Extended author information available on the last page of the article

1 Introduction

Social media website has become a platform and forum where users express emotions and opinions in diverse subjects such as politics, events, individual, products, dialogue systems and review ranking as well as summarization (Bharti et al. 2016). It has also become a popular platform for global interaction and idea discussion among users. Many firms have realized the necessity of analyzing the social media data in order to get the emotion of the customers regarding their products, which will, in turn, increase the quality of their products. The subjective and emotional language often requires a specific context in order to comprehend the meaning of what the user is discussing. Sarcasm, according to the Cambridge English dictionary is defined as 'the use of remarks that clearly mean the opposite of what one says, made in order to hurt someone's feelings or to criticize something in a humorous way' (Dictionary 2008). Similarly, Macmillan English dictionary defines Sarcasm *as* 'the use of remarks in saying or writing the reverse of one's motive in order to hurt someone's perception' (Dictionary and Rundell 2007). Moreover, sarcasm is a figurative language often used in verbal and written text form to communicate in microblogs, such as Twitter. In sarcasm sentiment, the negative emotion of people is communicated using a positive term in the text to reveal their sarcasm. Sarcasm exists in many kinds of structure and order such as verbal or written sarcasm. The verbal sarcasm that usually occurs in speech can also be referred to as spoken sarcasm. Features like pitch level and variation, speech time and tempo, as well as acoustic features (intensity, volume, and frequency), are found in verbal sarcasm. This kind of sarcasm uses tones and gestures like eye and hand movement to show their sarcastic features. In contrast, written sarcasm occurs in a medium such as official letter, email, social media, and product review. In the one hand, when sarcasm is used in communication, it becomes hard to efficiently identify by employing data mining approaches due to the differences in its implicit and explicit meanings in a sentence (Yee Liau and Pei Tan 2014). On the other hand, when sarcasm utterance is expressed in a textual data, it is difficult to be identified by a common person due to the absence of tune and gesture in the textual data (Bharti et al. 2016). Therefore, an efficient natural language processing (NLP) method for text classification in a sentence that possesses sarcastic attributes and properties is required to identify sarcasm (Yavanoglu et al. 2018).

Various authors have defined sarcasm in terms of NLP approaches. For instance, Yavanoglu et al. (2018) defined sarcasm identification as an activity of using NLP techniques to classify a word or sentence sequence that possesses sarcasm attributes and properties. They also referred to it as the system that learns and distinguishes between normal sentence and sarcasm within the semantic level in a sentence. The main objective of sarcasm identification in a sentence is sentiment classification. Thus, the machine-learning model is often employed for sarcasm identification due to its durability and competence to observe itself in conformity with the datasets and specifications. There are various areas that sarcasm identification has played critical roles. For instance, sarcasm identification experiment enhances the research on sentiment analysis. In such a case, emotion features serve as a bedrock for sentiment polarity identification and opinion mining classification. In addition, sarcasm identification enables companies to analyse feelings of customers regarding their products; this could improve the quality of their products (Saha et al. 2017). It is also helpful in the reduction of the wrong categorization of consumer's opinions towards issues, products, and services (Mukherjee and Bala 2017b). Moreover, sarcasm identification is useful in dialogue, system review ranking, and summarization in human–computer interaction application domains (Davidov et al. 2010).

Lately, few review and survey studies have been published on sarcasm identification in the social media (Wicana et al. 2017; Yavanoglu et al. 2018). For instance, Wicana et al. (2017) presented a machine learning-based review on sarcasm identification by explaining the most current used classification algorithm for sarcasm identification such as support vector machine, maximum entropy, winnow class, neural network, semantic, and statistics, among others. Similarly, Yavanoglu et al. (2018) presented a technical review on sarcasm detection algorithm and explained the most currently used algorithm for sarcasm identification. In addition, Joshi et al. (2017), carried out an in-depth survey on automatic identification of sarcasm and reported the comparison in the magnitude of the study such as the approaches, features employed, classification algorithm and the performance parameter, which is useful in the understanding of the latest trends in identifying sarcasm. According to their study, three discoveries have been identified since the history of sarcasm identification namely pattern extraction using semi-supervised identification, supervised learning with the use of hashtag and context usage above the target text.

However, there are inherent limitations with the current reviews mentioned above. Firstly, the above reviews have failed to provide a comprehensive review of the dataset employed for sarcasm identification. Secondly, none of the studies has provided an extensive and in-depth review of recent pre-processing techniques for sarcasm identification, though a pre-processing step is a key step in any classification problem. Furthermore, none of the reviews has been able to provide a comprehensive review on feature selection and representation schemes for sarcasm identification. Besides, a review of performance parameters for sarcasm identification was also omitted in the previous reviews. However, the aforementioned limitations that are found in the current reviews have motivated the authors for a thorough review and study on sarcasm identification approaches using textual data. Nonetheless, the investigation shows that no studies have been conducted on the comprehensive reviews on sarcasm identification in well-known databases such as Scopus, IEEE Explore, Web of Science, Association for Computing Machinery, Google Scholar, Science Direct and Springer. Hence, there is a need for a systematic study to find out the present state-of-the-art sarcasm identification in the social media.

Consequently, the aim of this review is to present an extensive review and analysis on identification of sarcasm on published articles starting from 2008 to 2019 by exploiting and critically reviewing sarcasm identification under the following perspectives: The datasets, feature usage, feature engineering techniques (feature extraction technique, feature selection techniques, and feature representation technique), classification algorithm and the performance parameters. Forty (40) academic literature were selected after an in-depth search from the six familiar academic databases to carry out the review. The purpose of conducting this review is to help scholars in carrying out research in the area of sarcasm identification by answering the under listed research questions:

Research Question 1 Are there annotated sarcastic datasets publicly available in the area of sarcasm identification using text classification methods?

Research Question 2 What are the most useful features for sarcasm identification by researchers and why?

Research Question 3 Which feature extraction techniques are mostly often employed in sarcasm classification methods and why?

Research Question 4 Which feature representation technique is mostly applied in sarcasm classification methods and why?

Research Question 5 Which feature selection techniques do researchers in sarcasm classification commonly embrace?

Research Question 6 Which of the text classification algorithm produces better accuracy and why?

Research Question 7 Which performance measures are most widely used to measure the performance of the classifiers in sarcasm classification?

Research Question 8 What are the directions for future research and challenges in the domain of sarcasm classification?

The major contributions of this systematic literature review to the current body of knowledge in sarcasm identification in the social media are:

- A comprehensive investigation of characteristics, types, strengths, and weaknesses of datasets for sarcasm identification in the social media textual data.
- An outline of effective approaches for sarcasm identification, in conjunction with the various features representation and extraction scheme for efficient algorithm development.
- A critical analysis of various data preparation (pre-processing) techniques and classification algorithms (classifier) for sarcasm identification.
- Identification of recent research challenges and suggestion of open research direction to tackle issues in sarcasm identification domain.

The remainder of this article is divided into six sections. Section 2 describes the approaches for sarcasm identification such as the lexicon-based, the rule-based, and the machine learning based. Section 3 discusses text classification stages and techniques. Section 4 explains the research methodology for this review. Section 5 provides an extensive review of the selected articles under five different phases such as datasets, feature sets, feature engineering techniques, text classification techniques, and performance measures. Section 6 presents the research challenges and the future research direction on sarcasm identification in the text classification domain. Finally, Sect. 7 provides the conclusion of the study by giving a summary of the review findings. The review structure is shown in Fig. 1 whereas the list of abbreviations used in this review with their full forms is shown in the "Appendix".

2 Sarcasm identification approaches

Researchers have carried out studies on sarcasm identification in textual data. Various studies approaches for automatic identification of sarcasm found in literature are lexicon-based (Riloff et al. 2013), rule-based NLP (Mukherjee and Bala 2017a), pattern-based (Bouazizi and Ohtsuki 2016), lexicon-based approach (Bharti et al. 2015), Corpus-based (Khodak et al. 2017), statistical-based approach (Reyes et al. 2013) and machine learning based (González-Ibánez et al. 2011). Recently, deep learning approach (Ghosh and Veale 2016; Mehndiratta et al. 2017) which is a new trend, has also gained considerable ground on sarcasm identification and few researchers have employed the approach. The detailed explanations of those approaches are presented in the subsections below.

2.1 Lexicon based approach

In lexicon-based approach, a bag-of-lexicon (comprising unigram, bigram, trigram. etc.) and phrases are used to recognize sarcasm in tweets. For instance, Riloff et al. (2013), utilized a bootstrapping method to construct two bags-of-lexicon that consist of unigram,

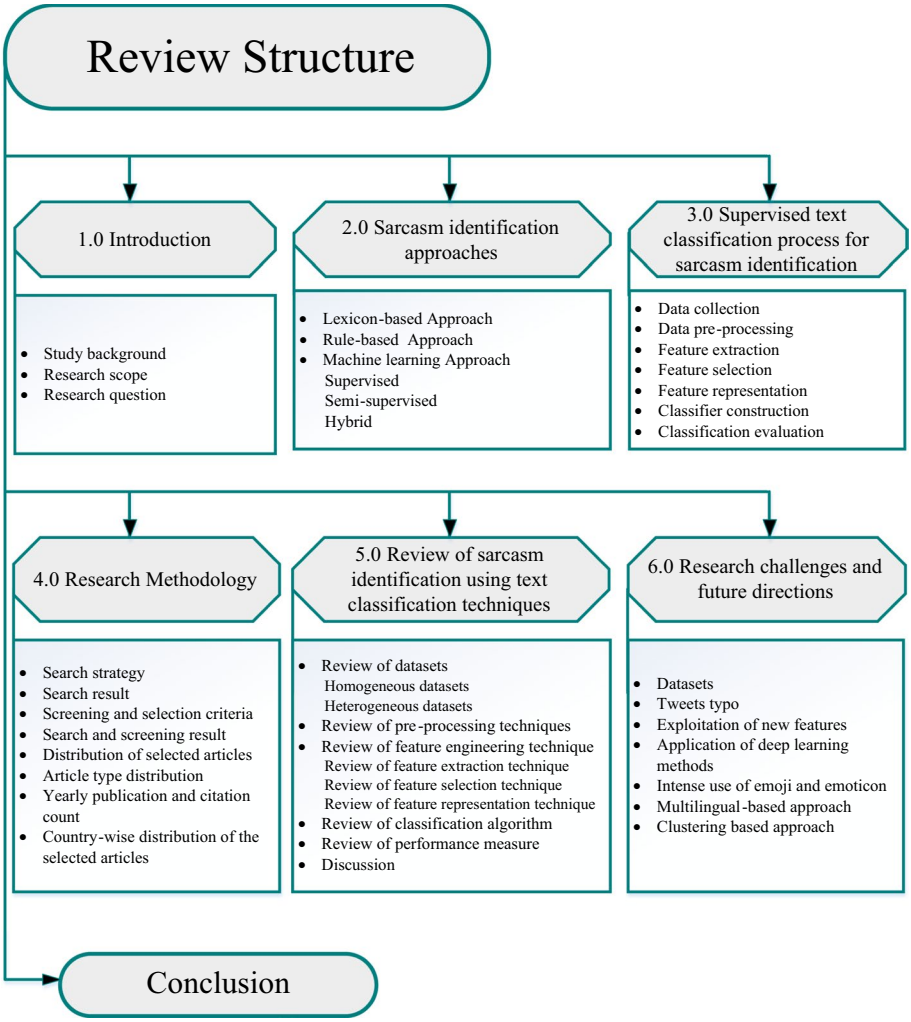

Fig. 1 Review structure

bigram and trigram phrases. Moreover, these phrases were employed for sarcasm identification in tweets, where the positive sentiment is used in a negative situation. Comparably, four bags-of-lexicon consisting of positive sentiment, negative sentiment, positive situation and negative situation have been developed (Bharti et al. 2015). However, they employed these phrases to recognize the occurrence of sarcasm as negative sentiment in a positive situation and positive sentiment in a negative situation.

2.2 The rule-based approach

In sarcasm identification, the rule-based approach is a problem finding method, which uses an object that relies on some specific principle or guideline. The rule-based approach uses

syntactic, semantic and stylistic properties of the sentence such as the pattern of phrase and lexical structure of sentence analysis in any language for sarcasm identification. Most researchers often employ this approach as a means of result comparison with the classifier being used. The semantic-based approach, one of the rule-based approaches, emphasizes more on the meaning of word use, its structure, structural relationship of the word and the contextual usage in the language (Liu 2012). The semantic-based model is the bedrock of the rule-based approach due to its effectiveness in nature. Accordingly, one of the studies that utilized this approach for sarcasm identification was presented by (Bharti et al. 2015). The study used Twitter dataset and the feature extraction techniques that comprise parsing, parts-of-speech tagging and parse tree to learn the semantic arrangement of a sentence. The study employed two algorithms to determine the diverse polarity sentiment in a tweet and the tweets that started with interjections. However, their result shows that the most sarcastic sentences begin with an interjection in a sentence. Similarly, Riloff et al. (2013) also presented a rule-based algorithm that searches for the occurrence of a negative situation and positive verb phrase in a sentence. The study utilized a well-structured iterative algorithm for the extraction of the negative situation phrase and carried out the experimental analysis with various sets of the rule.

2.3 Machine learning approach

This approach is one of the most applied approaches for sarcasm identification by researchers. This is because of its stability feature and its ability to observe itself in correspondence with a dataset and a given specification. Machine learning approaches deals with the creation of a prediction model using an intelligent method. The effect of pragmatic and lexical aspects in machine learning algorithm was studied in (González-Ibánez et al. 2011). The machine learning approach can be further be categorized into unsupervised learning, supervised learning, semi-supervised learning, structural and hybrid learning. A brief explanation of these approaches is given below.

2.3.1 Supervised learning

Among the machine learning algorithms, supervised learning is mostly used in sarcasm detection because of its ability to build a model by taking a labelled dataset as an input data (Mohri et al. 2012), and producing a labelled output data which helps in the construction of a descent model. This is made possible because the training datasets have already provided the result that is to be processed by the model. Supervised learning algorithm (like NB, DT, and LR) serves as the bedrock for other learning algorithms with similar precepts (Yavanoglu et al. 2018). The machine learning algorithm (such as SVM and LR) in addition with the sequential minimal optimization (SMO), was also employed to differentiate sarcasm from the polarity sentiment occurring in Twitter messages (González-Ibánez et al. 2011).

Furthermore, the popularity of the architecture of deep learning approaches has created an opportunity for researchers in this domain to conduct a study on the automatic identification of sarcasm. This form of learning consists of a subset of machine learning by employing neural network to automatically learn from large datasets (Nweke et al. 2018). A neural network is a learning algorithm that processes the features similar to the functioning of the nerve system in the human brain. In the neural network, each unit of the network has a connection to many other units, which can possess a summation function

that combines all its input value together. The neural network uses 0.0 and 1 real number value representation in terms of core and axon. Recently, Ghosh and Veale (2016) employed a deep neural network model to identify sarcasm occurrences on twitter datasets. In their work, they combined the algorithms that consist of a convolutional neural network, long short term memory (LSTM) network, and recursive support vector machine and got an impressive performance of the model over the baseline method for sarcasm detection system of F-score of 92% (Schifanella et al. 2016). Similarly, Joshi et al. (2016) in their study also used features based on the similarity of word embedding for sarcasm identification. The feature used in their study was enhanced in relation with the most congruent and incongruent word pair, which resulted in an improvement of the performance.

2.3.2 Semi-supervised learning

This form of machine learning algorithm is a mixture of supervised and unsupervised learning using a minimal quantity of annotated data and a vast number of unannotated data (Tsur et al. 2010). The presence of the unlabelled datasets, and the open access to the unlabelled datasets is the feature that differentiates the supervised learning from the semi-supervised learning. This type of learning approach was employed by Davidov et al. (2010) for automatic sarcasm identification using amazon product review datasets. In their study, a total number of 66,000 products and book reviews were collected and both syntactic and pattern-based features were extracted. The sentiment polarity of 1 to 5 was chosen on the training phase for each training data. The authors reported a promising performance of 77% precision and 83.1% recall on the evaluation phase.

2.3.3 Hybrid learning based

This approach is a mixture of two or more classifiers to form a new one. In other words, it refers to an ensemble classifier. A study that employed the approach is the learning of user-specific context presented by Amir et al. (2016), it uses a convolutional network to learn user embedding feature in conjunction with the utterance-based embedding feature. The resultant features formed a hybrid convolutional user embedding convolutional neural network (CUE-CNN) model in the domain of sarcasm detection and the result of the study produced a performance increment of 2% over single machine learning approaches for sarcasm identification.

3 Supervised text classification process for sarcasm identification

According to Nithya et al. (2012), supervised text classification is a classification that makes use of labelled training datasets of the text to learn and build a text classifier that can be used to automatically classify the unlabelled test sets. Human observers are often used to perform text categorization nowadays, however, these are deemed incompetent owing to the huge amount of files, email messages and web addresses that are being saved in a folder every day. Moreover, manual categorization is usually slow and costly to maintain. In addition, inconsistency is another limitation inherent in manual categorization. The above-identified limitations have shifted the text classification from a manual to an automated base. Several techniques exist in automated text classification such as supervised, semi-supervised and unsupervised text classification. However, the supervised approach is

most globally used as it has the ability to build a model using labelled data as an input data (Mujtaba et al. 2017; Yavanoglu et al. 2018). Supervised text classification experimental process consists of six main steps as explained in the subsequent sections.

3.1 Data collection

Data collection phase comes first in any text classification process. The collection of dataset is in relation to the domain the study is considering. For example, when a study seeks to detect sarcasm on Twitter, then the Twitter data is collected. When a study seeks to analyse the disaster response and recovery through sentiment analysis, then the disaster-related data is collected in the social media. In any case, once the raw data is collected, the next phase of the classification is to pre-process the data before the actual analysis can be carried out on the dataset.

3.2 Data pre-processing

Raw data collected during the data collection phase contains a lot of noisy information and requires cleaning. The purpose of cleaning is to eliminate the noise from the data before some knowledge or features can be extracted from it. In addition, duplicate data are also removed during pre-processing stage especially the social media data (Eke et al. 2019). Data pre-processing is referred to as the data preparation phase, where the training and testing datasets are prepared. In the training sets, Twitter datasets are labelled as either sarcastic or non-sarcastic and are required to train the model, whereas the testing datasets are not labelled since it is mainly used for model evaluation. The pre-processing stage mainly seeks to remove unnecessary characters or sequences, which have no value to the sentiment classification. In this phase, the collected data will first undergo a tokenization process also called automatic filtering. This is purposefully performed to remove retweets, duplicates, stop words, punctuations, numerals, tweets written in other languages and tweets with the only URL. At the end of the filtering stage, parts of speech (POS) tagging and stemming is then applied to the remaining tweets in order to covert the text to its original form.

3.3 Feature extraction

Feature extraction is the third stage in the supervised learning approach with regards to the text classification task. It is a technique used to reduce the number of resources required for the description of the dataset by transforming the input data into a set of features. The feature consists of linguistic, pragmatic, emotional, psychological, hyperbolic features, among others. Section 5.3 provides more explanation on these features. The most commonly used feature engineering techniques are Bag-of-words and N-gram techniques. The Bag-of-words model is a text classification technique that uses the frequency of each word as a feature for classification. The bag-of-word technique is one of the widely used techniques for document representation in information retrieval for some years now and as a tool for feature generation (Salton and McGill 1986; Yavanoglu et al. 2018). However, in the N-gram technique, n stands for the number of word features. For example, when the value of n is 1, the feature is called unigram, when n is 2, it is called bigram, and when n is 3, it is called trigram, and so on. Simplicity and scalability are one of the choices of using this technique over the bag-of-words model (Yavanoglu et al. 2018).

3.4 Feature selection

The whole feature sets extracted from the datasets contain irrelevant features that may limit the prediction result during the classification stage. For instance, drawbacks in during the text classification due to the immaterial feature content are a reduction in the accuracy, a problem in generating a result, a decrease in the classification process and a difficulty in storage and retrieval of information. Hence, there is a need for a feature selection technique to choose the most discriminant feature subsets from the extracted feature sets for better prediction (Guyon and Elisseeff 2003). A thorough understanding of the aspect of the datasets that are relevant for the prediction that is to be carried out is needed. Feature selection technique can be sub-divided into wrapper, filter-based and embedding techniques (Guyon and Elisseeff 2003). Among these three categories, the filter-based technique is widely employed (Yang and Pedersen 1997). The filter-based technique uses statistical means to allocate a score to each feature and the selection and rejection of the feature are determined by the score. Chi square (χ^2) and information gain (IG) are common examples of feature selection filter-based technique. However, the wrapper-based approach uses the query technique for the best feature selection from the different combination and performs an evaluation using other combinations, whereas the embedding method studies the essential features on the course of building the model.

3.5 Feature representation

In text classification, the feature extracted is converted into a numerical value during the feature representation step (Salton and Buckley 1988). Feature representation technique is categorized into term frequency (TF), binary representation (BR), and term frequency with inverse document frequency (TFIDF) (Debole and Sebastiani 2004). In the TF representation, the value of the feature signifies the total occurrences of the feature in the document (Ramos 2003). However, in the BR technique, the feature value 0 or 1 is used for representation where value 1 indicates the presence of the feature in the document and value 0 signifies the absence of the feature in the document (Salton and Buckley 1988). In IF-IDF representation, the frequency of the text in a particular document is calculated and the result is compared with the inverse portion of the frequency of the word in the whole document. It is effective in matching a word in a query to documents that are important to the query (Ramos 2003).

3.6 Classifier construction

At this phase, the classification model is created on the training datasets by utilizing the machine-learning algorithm. The created model has the ability to classify the unlabelled data as sarcastic or non-sarcastic. Several algorithms have been implemented for the purpose of sarcasm identification. Few of the algorithms used in the selected studies consist of Naïve Bayes (NB), support vector machine (SVM), decision tree (DT), random forest (RF), linear regression (LR) and artificial neural network (ANN) (Yang 1999). These algorithms are described in the subsection below.

3.6.1 Naïve Bayes

Naïve Bayes (NB) is a classification algorithm that uses a probabilistic model to predict how data is obtained within a given class. It is a machine learning algorithm that performs a statistical analysis of numerical data (Sahami et al. 1998). It uses a labelled set of data as input data to calculate the parameter of the generative model. It is one of the simplest learning classifiers that assumes that all features do not depend on each other in a given class context (McCallum and Nigam 1998). Moreover, NB is one of the fastest classifiers that perform well when Bag-of-words techniques are used in text representation (Rennie et al. 2003).

3.6.2 Decision tree

Decision tree (DT) is a core algorithm employed in data mining for classification as well as for prediction. It is an induced learning algorithm that is centered on the instances, it concentrates on the classification rule that displays a decision tree deduced from a group of disorder to an irregular instance (Dai et al. 2016). The tree consists of leaf node, path, decision node and edges (Quinlan 1990). DT is a classifier that is represented in the form of a flow-chart tree structure, in which a core node represents the attribute test, each branch denotes a test result and each leaf node denotes a class. Thus, the whole tree tallies to a collection of a disjunctive representation rule (van der Aalst 2001). DT is employed to train instance classification, which can classify instance based on the definite attribute occurrence of the value sets. Over-fitting problem is one of the limitations inherent in a decision tree classifier. This is due to its capability of fitting every category of data along with the noise that can extremely influence its performance. Notwithstanding, this problem can be overcome by employing multiple classifier model such as the random forest in which different trees are designed and trained by dividing the training set, and the final predictions are combined over the tree.

3.6.3 Random forest

Random forest (RF) is an ensemble classification that uses sub-training sets to build a decision tree classifier. As such, DT classifies each of the input vectors in a forest and the most predicted classifier is selected. Random forest solves the over-fitting problem and it produces better prediction compared to a single decision tree (Liaw and Wiener 2002; Fernández-Delgado et al. 2014).

3.6.4 Support vector machine

Support vector machine (SVM) is a supervised learning algorithm that builds a classification model using the learning theory of statistics. The classification task requires the separation of the data into the training set and the test set. However, it uses the training set to build a model that predicts the target value of data, giving only the test data attributes (Hsu et al. 2003). In a support vector machine, a hyper-plane, also known as a support vector is used to separate the two-class data points by reducing the space between them with the help of training sets (Cristianini and Shawe-Taylor 2000). Many

applications such as sarcasm detections, image classification, and bioinformatics have been successfully carried out using the SVM classifier (Fernández-Delgado et al. 2014).

3.6.5 Maximum entropy classifier

The classifier that depends on the maximum entropy chooses from all the models the classifier with the highest entropy that fits the training data. This model does not presume the conditionally independent feature and as such, has a lesser restrictive model than other classifiers. Maximum entropy classifier has an optimization problem that requires handling in order to calculate the parameters of the model. Consequently, it requires more time for training compared to other classifiers like NB classifier (Mukherjee and Bala 2017a).

3.6.6 Artificial neural network

A neural network is a learning algorithm that possesses the features similar to the functioning of the nerve system in the human brain. An artificial neural network comprises of three distinct layers; input, hidden and output layer. While the input and hidden layers consist of numerous nodes, the output layer is made up of just one node. In the neural network, each unit of the network has a connection to any other units, which can possess a summation function that combines all its input value together. The hidden layer is designed for input processing and it connects to the output layer that garbage out the output values. The Neural network uses 0.0 and 1 real number value representation in terms of core and axon (Yavanoglu et al. 2018). According to Yao (1999), learning in artificial network is categorized into unsupervised, supervised and reinforcement learning. The unsupervised approach centres on the relationship that exists among the input data. In that regard, there is unavailability of "correct output" information for the learning. In supervised approach, the learning is based on comparison between the actual input and the target output of artificial neural network, in order to reduce the error function that exists between them. In so doing, the gradient decent-based optimization such as back propagation is employed to iteratively regulate the connection weight in order to reduce the error. Reinforcement learning on the other hand is a special case of supervised approach that provides information on the correctness of the actual output. In that case, there is no knowledge of the precise desired output. In an artificial neural network, learning rule is utilized for weight modification on each input pattern and the most commonly used rule is Delta rule (He and Xu 2010).

3.7 Classification evaluation

In the evaluation phase, the formulated classifier predicts the class of unlabelled text (sarcastic or non-sarcastic) using the training data sets. The classifier accuracy can be estimated by evaluating.

- The instance accurately classified in the correct class [true positive (TruPos)].
- The instance accurately predicted in the correct classes that are not members of the class [true negative (TruNeg)].
- The instances that were either inaccurately predicted to the particular class [false positive (FalsPos)] or that were not predicted as the instance of the class [false negative (FalsNeg)]. These four members consist of the confusion matrix for the binary classification as indi-

Table 1 Confusion matrix

	Actual instance	
	Yes	No
Predicted instance		
Yes	Tru Pos	Fals Neg
No	Fals Pos	Tru Neg

cated in Table 1. Various performance parameters have been employed for the evaluation of the model performance. The most commonly used measure for text classifications is accuracy, F-measure, precisions and recalls. They are briefly described below.

3.7.1 Accuracy (Acc)

The accuracy provides the percentage ratio of the predicted instance. It measures overall correctly classified instance. It is defined as

$$Acc = \frac{TruPos + TruNeg}{TruPos + TruNeg + TruNeg + TruNeg}. \tag{1}$$

3.7.2 Precision (Pre)

Precision is a computation ratio of *true positive* over positive result

$$\Pr e = \frac{TruPos}{TruPos + FalsPos}. \tag{2}$$

3.7.3 Recall (Rec)

Recall is the proportion of actual positives, which are predicted positive. It computationally represents the ratio of *true positive* against all the *true* result.

$$\mathrm{Re}c = \frac{TruPos}{TruPos + FalsNeg}. \tag{3}$$

3.7.4 F-measure (F-m)

F-measure represents the harmonic mean of precision and recall particularly when there is severe equality of *false positive* and *false negative*. The standard F-M is F1, which gives precision and recall equal importance.

$$F - m = 2 \times \frac{Pre \times Rec}{Pre + Rec}. \tag{4}$$

4 Research methodology

The aim of this study is to conduct a review of sarcasm identification classification in textual data. This section provides the research methodology adopted for the study. The study adopted systematic literature review (SLR), a guideline established by Kitchenham et al. (2009) for the computer technology field. The guideline consists of four major different phases namely planning for the study, primary study search and selection, data acquisition, and analysing of data. Initially, the planning phase establishes the problem statement, formulates the objectives, research questions and review protocol for the study (covered in Sect. 1). The search process consists of the search strategy, the search query, the selection criteria, and the search keyword on the screened study (to be explained in Sect. 4.2). The data collection phase applies the data extraction strategy on the retrieved study as explained in Sect. 4.3. Finally, the data analysis stage that combines the systematic review involves data synthesis and extensive analysis of the selected studies as explained in Sect. 5. The process of the review methodology is presented in Fig. 2.

4.1 Search strategy for the study identification

This study carried out an electronic search from seven major academic databases viz ACM Digital library, Web of Science, IEEE Explore, Science Direct, Springer, Google Scholar, and Scopus. The study considers the articles published from 2008 to 2019. In the search strategy, different suitable keywords were defined to search the literature on "Sarcasm identification on social platform" from the chosen databases. The search keywords used are sarcasm identification, sarcasm detection, sarcastic text, sarcastic sentence, sarcasm in microblog, sarcasm, sarcastic, sarcasm in a social platform, sarcasm in social media, and sarcasm in twitter. The synonyms of the formulated keywords were used to create additional keywords for search such as cynicism in a social platform, mockery remark in microblog, and satire utterance in social media. All the articles written in English language were investigated irrespective of the language used for the data analysis. The article type and language screening were employed. Finally, an extensive full-text evaluation review was carried out on the selected articles for suitable studies based on datasets, feature engineering, classification, and evaluation.

4.2 Search result

In this section, the queries based on the search keywords were applied to the entire seven-selected databases to fetch the academic articles. Thus, a total number of 51,069 articles were gathered. Table 2 shows the thorough search result from each academic database. The duplicate copies obtained across different databases were removed and only the distinct copies were retained and saved in the endnote.

4.3 Screening and selection criteria

After removing the similar studies occurring in more than one database, the total number of forty-two (42) articles was further analyzed by reading the title, abstract, and keywords to find out if these retrieved articles were obviously relevant to the purpose of carrying out

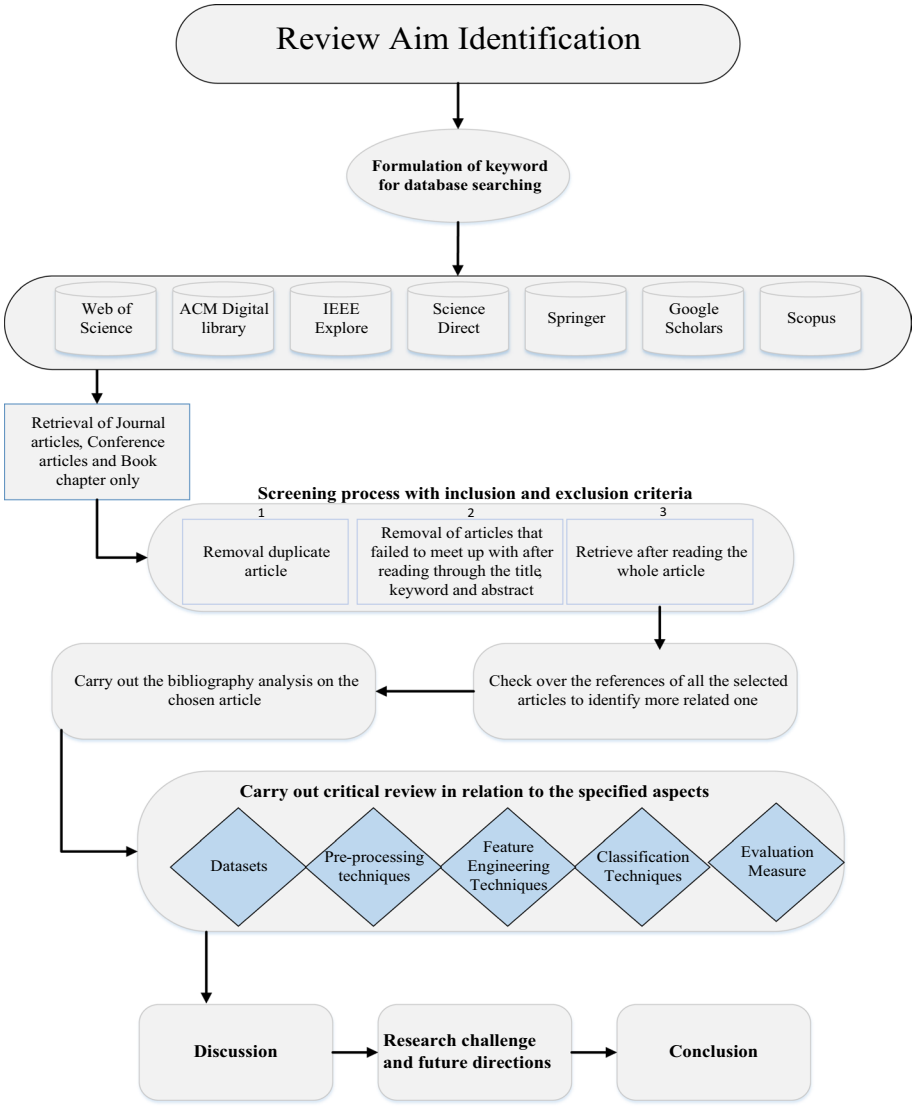

Fig. 2 Methodology process

the systematic review. This process is called screening stage 1. The output of this screening stage produced forty-two (42) articles, which were finally read intensely to see whether they tally with the inclusion criteria. This is called the screening stage 2. The output of this stage of the screening filtered thirty-six (36) articles. Lastly, the references of the thirty-six (36) articles selected were scanned to find some more related articles that conform to the inclusion criteria. This is called screening stage 3, and the output of the scanned references produced additional four (4) new articles. Therefore, a total number of forty (40) articles was selected for detailed analysis for the seven major academic databases, as indicated in Table 2. However, the selected articles were extensively reviewed under the following

Table 2 Search and screening result from the 7 databases

Database	Queries result	Screening 1	Screening 2	Screening 3
Web of science	616	3	2	2
ACM digital library	33,884	8	7	7
IEEE explore	39	6	6	8
Science direct	3331	5	3	4
Springer	742	4	4	4
Google scholar	12,300	11	10	11
Scopus	157	5	4	4
	51,069	42	36	40

consideration: (1) dataset for the study, (2) the pre-processing techniques (3) feature engineering techniques (4) classification techniques and (5) performance metrics (section five gives the detail discussion). The selection criteria is shown in Table 3.

The academic database wisely distribution of the forty (40) selected articles for the study is shown in Fig. 3. In the 40 articles, 2 were selected from the web of science, 7 from ACM digital library, 8 from IEEE Explore, 4 from Science Direct, 4 from Springer, 11 from Google scholar, and 4 from Scopus.

Figure 4 represents the selected studies distribution according to the type of article used for the study. The figure shows that 25 articles out of the 40 selected articles are conference proceeding articles, 12 articles are journal articles, and 1 article is a book chapter.

The yearly distribution publication count and the yearly citation count of the articles are shown in Fig. 5. In the chart, the vertical axis represents the number of articles published in years and the citation count obtained on the article that year, whereas the horizontal axis represents the year of publication. The optimum number of publication on the selected articles was attained in the year 2013, followed by 2010, 2011, 2015, etc. There is a decreasing trend in publication and citation count between 2017, 2018 and 2019. The figure also shows that there is no publication identified in the year 2008, 2009, and 2012 on the targeted topic.

Table 3 Selection criteria

S. no.	Inclusion criteria
1	The article must have been published in the English language
2	The article must have been published between 2008–2019
3	The article must be either a journal article or conference article or book chapter
4	The article must use machine learning for classification such as supervised or unsupervised or deep learning or semi-supervised learning
5	The article must use social media dataset
6	The article must use feature set, feature engineering techniques, propose new classification or clustering technique or use the existing classification or clustering technique
	Exclusion criteria
1	Studies that are not purposefully for sarcasm identification
2	Studies that do not meet up with any of the stated inclusion criteria

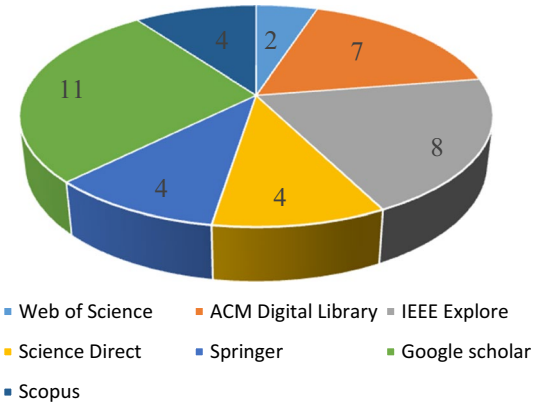

Fig. 3 Distribution of selected articles

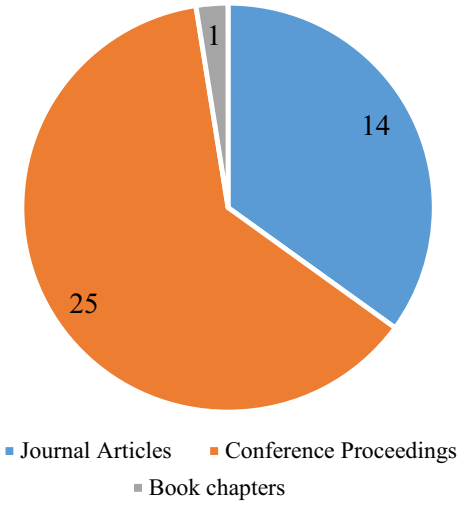

Fig. 4 Article type distribution

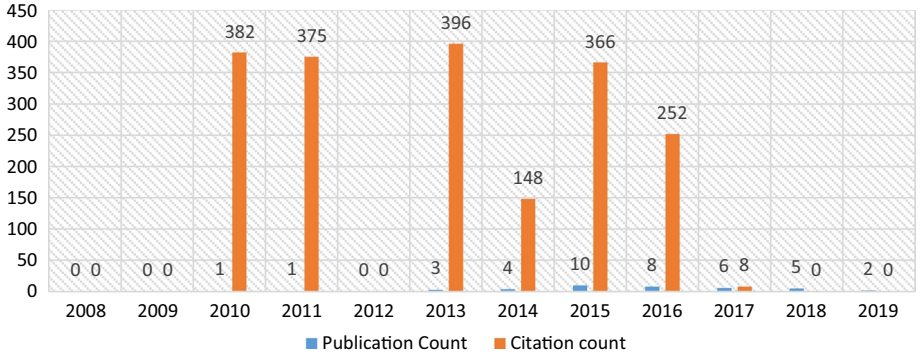

Fig. 5 Yearly publication and citation count

The country-wise distribution of the selected articles is also shown in Fig. 6. It is obvious from Fig. 6 that the largest number of the selected articles on the topic was published from the USA, succeeded by India, Netherland, Indonesia, Japan, Portugal, China and UK, Philippine, Sweden, Australia, Ireland, Tunisia, France, Slovenia, and Vietnam.

5 Review of sarcasm identification using text classification technique

In this section, a critical review of the selected primary study on various aspects was carried out. The aspects consist of datasets usage, pre-processing techniques, feature engineering techniques, the modelling approach, and performance metrics. The section is divided into various subsections. In Sect. 5.1, the reviews of the various datasets used for sarcasm identification were presented. Section 5.2 presents a review of various pre-processing techniques used for sarcasm detection. Section 5.3 presents a review of feature engineering techniques used for sarcasm identification. In Sect. 5.4, a review of different modelling approaches used for sarcasm identification was presented. Lastly, Sect. 5.5 gives a review of various performance metrics used for classification performance evaluation for sarcasm detection.

5.1 Review of datasets for sarcasm identification

The sarcasm identification dataset is an important component of the sarcasm classification task. However, such dataset is worthless on its own except some features or useful knowledge are extracted from it. Related studies on sarcasm text classification showed that authors collected primary data using social media and employed two main annotation strategies such as distant supervision via hashtag (Abercrombie and Hovy 2016) and manual annotation strategy (Riloff et al. 2013). The first stage in sarcasm identification experiment is the collection of data to be utilized for building the classification model. The analysis of the selected studies for sarcasm identification shows that datasets can be broadly categorized into homogeneous and heterogeneous data. These data categorizations review are

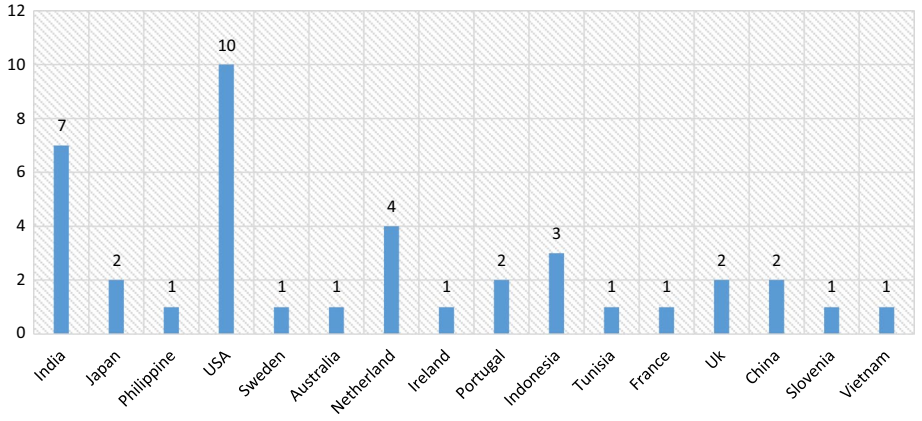

Fig. 6 Country-wise distribution of the selected article

explained below while the strengths and weaknesses of deploying the datasets for sarcasm identification are shown in Table 4.

5.1.1 Homogeneous data

In homogeneous data, the studies utilized only one type of dataset which is majorly from the Twitter platform. For instance, a study on 'Sentence level sarcasm detection in English and Filipino' that was carried out by (Samonte et al. 2018) utilized only twitter datasets. The researchers collected a total number of 12,000 tweets consisting of 6000 Tagalog and 6000 English tweets. The authors employed datasets on topics such as transportation, government, politics, social media, and weather. In the study, face pager API was utilized for the collection of data from Twitter. The parameters on face pager were set accordingly such as the result type (result_type); that specifies the preferred result by the users (i.e. popular, recent or mixture of both), the count; that specifies the maximum number of tweets to be retrieved (usually 200 maximum), and the language type; that specifies the type of language of the returned tweets. However, similar parameters settings were used for both English and Tangalog tweets collection except in the language specification, in which *tl (for Tangalog)* was used on Tangalog dataset. Thus, the study indicated that the nature of the datasets (balanced or Imbalanced) has a great influence on the model's prediction in terms of the accuracy for sarcasm. In addition, (Kumar and Harish 2018) used a content-based feature selection technique to build a classification model for sarcasm identification. The study utilized amazon product review datasets created by the study carried out in (Filatova 2012) and sourced from a crowd sourcing platform-*Mechanical Turk*. A total number of 1254 Amazon products reviews, consisting of 437 reviews (sarcastic) and 817 reviews (non-sarcastic) were used for the classification experiment. Interestingly, the datasets were structured using a star rating (ranging from 1 to 5) and review comments written in English language. In another study, Zhang et al. (2016) utilized twitter datasets for sarcasm identification using a deep neural network. The tweets datasets were obtained by using twitter-streaming API with sarcasm hashtag (#sarcasm) and not hashtags (#Not) keyword. The study adopted the datasets obtained by (Rajadesingan et al. 2015b), in which a total number of 9104 tweets annotated by the author of the tweets was used for the experiment. In this regards, similar tweets IDs provided by them were used to stream the corpus. Similarly, the contextual tweets were obtained by employing Twitter API in each tweet. However, the hashtag for sarcasm and Not (#sarcasm and #Not) were removed on the historical tweet to prevent the use of explicit clue for sarcasm prediction. Furthermore, the author noted that the use of both balanced and imbalanced datasets was modelled and the experimental result shows that the imbalanced dataset accuracies are greater than the balanced counterparts with the conflicting value of the F-measure. Therefore, imbalanced data create biases in sarcasm identification and performances of the model.

5.1.2 Heterogeneous data

The dataset used here for identification of sarcasm is obtained from various social media and other platforms such as Instagram, Amazon, Tumblr as well as product reviews from electronic commerce in order to improve the robustness and generalization of the sarcasm identification model. For instance, Schifanella et al. (2016) utilized dataset obtained from Twitter, Tumblr and Instagram for sarcasm detection in the multimodal social platform which comprises of text and image datasets. In a previous work (Liu et al. 2014),

Table 4 Dataset and volume used on the selected studies

Data type	Data sources	Strengths	Weaknesses	References	Number of Studies
Homogeneous data	Twitter or Product Review	Management of the data collection process is easier and cost effective as the datasets are from a single entity. Furthermore, Twitter provides a rich source of API for data collection	It is challenging to provide high generalization for data from one source for sarcasm identification as it involves varieties of applications	González-Ibánez et al. (2011), Lunando and Purwarianti (2013), Liebrecht et al. (2013), Riloff et al. (2013), Barbieri et al. (2014), Ptáček et al. (2014), Altrabsheh et al. (2015), Bharti et al. (2015), Bouazizi and Ohtsuki (2015a, b), Fersini et al. (2015), Ghosh et al. (2015), Khattri et al. (2015), Kunneman et al. (2015), Rajadesingan et al. (2015a), Wang et al. (2015), Amir et al. (2016), Bharti et al. (2016), Bouazizi and Ohtsuki (2016), Ghosh and Veale (2016), Ling and Klinger (2016), Sulis et al. (2016), Al-Ghadhban et al. (2017), Bharti et al. (2017), Manohar and Kulkarni (2017), Mukherjee and Bala (2017a, b), Ranjan et al. (2017), Abulaish and Kamal (2018), Kumar and Harish (2018), Manjusha and Raseek (2018), Samonte et al. (2018), Sreelakshmi and Rafeeque (2018), Kumar et al. (2019) and Suhaimin et al. (2019)	36

Table 4 (continued)

Data type	Data sources	Strengths	Weaknesses	References	Number of Studies
Heterogeneous data	Twitter, Amazon, Instagram, Tumblr, and Product review	The fusion of data from multiple sources helps to improve generalization, sarcasm identification model reliability, robustness and performance result	Aggregation of data from various sources may increase computation complexity, and lead to high computation burden. In addition, it is difficult to fuse a large amount of dataset from multiple data sources	Davidov et al. (2010), Liu et al. (2014), Schifanella et al. (2016) and Dharwal et al. (2017)	4

the researchers evaluated their model by employing two corpora (English and Chinese) sarcasm feature. However, the English sarcasm verification was carried out in the first corpus, which is content of news articles sets adopted from Davidov et al. (2010), the Twitter datasets used by Reyes et al. (2012), and Amazon datasets provided by Burfoot and Baldwin (2009). Then, the second corpus, which was used to verify Chinese sarcasm features also consisted of three different datasets obtained from Sina Weibo, Tencent Weibo and Netease BBC, to crawl various topical comments. Invariably, the heterogeneous dataset employed in this study is highly imbalanced. Consequently, Area under curve (AUC) performance measure was employed for performance evaluation as it has been proven successful for providing better performance measure for imbalanced dataset compared to F-score by using true positive rate instead of precision. Furthermore, Davidov et al. (2010) study focused on sarcasm identification that deployed two multimodal datasets. In this study, the datasets used consists of tweets (5.9 million tweets) and Amazon product review datasets (66,000 product reviews), which were adopted from Tsur et al. (2010). The tweets data was streamed using #sarcasm hashtag included by the tweeter. However, there is inconsistent in the use of the hashtag since it is not known to all the users, hence, most tweeters do not explicitly apply the hashtag for tagging the sarcastic tweets. To this end, the tweets that included hashtag annotation can be regarded as the 'Secondary gold standard for the detection of sarcastic tweets'. Still, in this study, the Amazon product review consisted of 120 products. The corpus is the content of different books and electronic products reviews. In contrast with the tweets, amazon products datasets are longer in size, as some of the review sentences contained about 2000 words. Interestingly, the sentence structure and grammar in the product review are better than the tweets datasets.

Table 4 outlines the data types, sources, strengths, and weaknesses of the data utilized for sarcasm identification.

5.2 Review of pre-processing techniques for sarcasm identification

Pre-processing of social media data is necessary because of the irregular and informal form of data acquired. The purpose of pre-processing is to eliminate some problems inherent in such texts like a misuse of letter, use of acronyms, poor grammatical sentence and unnecessary repetition (Cotelo et al. 2015). In the pre-processing stage, meaningless data from the acquired dataset are removed in order to enhance the performance of the classification model. The pre-processing techniques that are mostly used in sarcasm identification research according to the previous literature include removal of stop word, empty space, punctuations, special symbols, conversion of uppercase letters to lower case, stemming, tokenization, POS tagging, lemmatization, removal of URLs and hashtags. Thus, the efficiency of this pre-processing techniques are reported in various studies under consideration. In recent studies, Al-Ghadhban et al. (2017) and Samonte et al. (2018) tested the impacts of inclusion or removal of URL, user mentions and stops word in the textual data for sarcasm detection in twitter. The experimental result showed that their removal enhances classification accuracy than when they are present. Some researchers in their studies Ghosh et al. (2015), Dharwal et al. (2017) and Abulaish and Kamal (2018) illustrated the application of stemming, tokenization and conversion of upper case letters to lower case for pre-processing tasks for sarcasm identification. These studies reported that the application of such pre-processing techniques produced a better performance in classification when compared with other studies. A couple of scholars (Altrabsheh et al. 2015; Abulaish and Kamal 2018) have also tested the removal of the white space character,

punctuation marks, numbers, and emoticon. Their reports showed the effectiveness of applying these pre-processing techniques for improved classification tasks. Nonetheless, Kunneman et al. (2015) tested the usage of punctuation marks as a feature for modelling in their study on 'Signalling sarcasm from hyperbole to hashtag'. The result of their experiment showed a better performance in classification when punctuation marks are present than when they were removed. Therefore, we can conclude that researchers should test the performance of the various technique of pre-processing on the sarcastic corpus to check the accuracy of the algorithm in classification. The summary of the pre-processing techniques applied in the selected studies is illustrated in Table 5. The analysis from Table 5 shows that many studies made use of basic pre-processing techniques, which revealed the effectiveness of the pre-processing in attaining a better accuracy in the classification task.

5.3 Review of feature engineering techniques for sarcasm identification

Feature engineering is one of the major steps in any classification problem. Three stages are involved in feature engineering stages; they are feature extraction, feature representation and feature selection (Mujtaba et al. 2018). The output of the feature engineering stage is in the form of the feature vectors (in numerical form), which serves as an input to the learning algorithm (SVM, RF, DT, etc.) for classification model construction and validation. The detailed explanation of these stages was given in Sect. 3 and the review is presented in the subsequent subsection.

5.3.1 Review of feature extraction techniques for sarcasm identification

In sarcasm identification, feature extraction is the process of extracting relevant and discriminant information from the sarcastic dataset, which will help in the training of the model for sarcasm identification. The review of the selected studies showed that the semantic properties of the sentence features were used in most studies; researchers also utilized automatic feature extraction technique to extract content-based and linguistic features. This was carried out by using the algorithm and various statistical methods. The content-based feature extraction technique consists of Bag of the word (BoW) (da Silva et al. 2014), word to vector (word2vec) (Lee et al. 2018) and n-gram (Sintsova and Pu 2016) technique. As revealed in Table 8.

Table 6, most studies utilized N-gram feature extraction technique on the selected studies. For instance, some authors (González-Ibánez et al. 2011; Rajadesingan et al. 2015a; Kumar and Harish 2018) utilized n-gram feature extraction technique for sarcasm detection and reported that n-gram technique is useful in extracting lexical features. One of the motivations of the n-gram model usage by the researcher is due to its simplicity and scalability (the matching scale of all the enormous sample datasets) properties. In another study Suhaimin et al. (2017), on sarcasm detection in the bilingual text, various NLP techniques were used to extract the combination of various features such as lexical, pragmatic, syntactic, prosodic and idiosyncratic. These features were trained using a non-linear SVM algorithm. However, the result shows that NLP selected features outperformed the baseline features such as bag-of-words, which demonstrated a better performance of the proposed method. Furthermore, lexicon sentiment based feature and pragmatic features (emoticons and user mentions) were extracted in a study by González-Ibánez et al. (2011) for sarcasm identification. The experimental analysis showed that the combination of such features

Table 5 Pre-processing techniques used in the selected studies

Pre-processing techniques	Number of studies	References
Removal of Twitter user mentions, URL, hashtag, duplicates, quotes, elongation, punctuation marks, retweet symbols, less than 3 or 4 words, neutral tweets, manual labeling and stop words	10	Davidov et al. (2010), González-Ibánez et al. (2011), Fersini et al. (2015), Rajadesingan et al. (2015a), Schifanella et al. (2016), Zhang et al. (2016), Bharti et al. (2017), Mukherjee and Bala (2017a, b) and Sreelakshmi and Rafeeque (2018)
Conversion of numeric characters into alphabets, lower case, removal of local repetition, punctuation marks, blank spaces, special characters, stop words and digits	3	Lunando and Purwarianti (2013), Altrabsheh et al. (2015) and Kumar and Harish (2018)
Tokenization, stripped with a punctuation mark, retain capital letters, part of speech tagging, stemming, stop word removal, conversion to capital letters, upper to lower case conversion, stemming, removal of URL and user mentions	9	Liebrecht et al. (2013), Riloff et al. (2013), Barbieri et al. (2014), Ptáček et al. (2014), Ghosh et al. (2015), Khattri et al. (2015), Wang et al. (2015), Dharwal et al. (2017) and Ranjan et al. (2017)
Removal of a retweet, hashtag, irrelevant tweet, emoji, links, lemmatization, tokenization, acronyms and URL removal, part of speech tagging (POS)	4	Bouazizi and Ohtsuki (2016), Ling and Klinger (2016), Al-Ghadhban et al. (2017) and Samonte et al. (2018)
Removal of hashtag, unwanted space using a regular expression, replacements of emoticon and acronyms using dictionaries, tokenization, stop word removal.	1	Manjusha and Raseek (2018)
Tokenization, removal of URLs, @mention, retweets, hashtags, ampersands, and extra white space, upper to lower case, double quotes, 1emoticons, numbers, and dots	1	Abulaish and Kamal (2018)
C1eaning, instance selection, normalization, transformation, POS tagging, tokenization	1	Manohar and Kulkarni (2017)
Tokenization, (punctuation, emoticons, and capitalization information were kept), removal of less than 3 letter word	1	Kunneman et al. (2015)
Removal of social media markers such as profile references retweets and hashtag, parsing, and splitting of multiple sentences using the Stanford splitter	1	Ghosh and Veale (2016)
Tokenization, spell checking and stop word removal	1	Suhaimin et al. (2019)

Table 5 (continued)

Pre-processing techniques	Number of studies	References
URLs, @mention, hashtag and numbers in tweets are replaced with placeholder	1	Kumar et al. (2019)
The pre-processing technique that was used was not mentioned	6	Liu et al. (2014), Bharti et al. (2015), Bouazizi and Ohtsuki (2015a, b), Amir et al. (2016) and Bharti et al. (2016)

Table 6 Feature extraction techniques used in the selected studies

Feature extraction techniques	References
Punctuation and Pattern based feature	Davidov et al. (2010)
N-gram, lexical and pragmatic features	González-Ibánez et al. (2011)
N-gram, sentiment polarity and interjection word	Lunando and Purwarianti (2013)
Sentiment and pattern	Liebrecht et al. (2013)
N-gram, pragmatic and pattern	Riloff et al. (2013)
POS tagging and sentiment	Ptáček et al. (2014)
N-gram, parts of speech, pragmatics and pattern features	Barbieri et al. (2014)
Punctuation symbols, lexical features and syntactic features	Liu et al. (2014)
Sentiment-based, lexical and punctuation features	Bouazizi and Ohtsuki (2015b)
Bag of Word, pragmatics and Parts of Speech	Fersini et al. (2015)
N-gram and sentiment	Khattri et al. (2015)
N-gram and pragmatics	Ghosh et al. (2015)
N-gram	Rajadesingan et al. (2015b)
POS tagging	Bharti et al. (2015)
Sentiment, punctuation, syntactic and pattern	Bouazizi and Ohtsuki (2015a)
N-gram, pragmatics and polarity label	Altrabsheh et al. (2015)
Sentiment-based feature	Wang et al. (2015)
N-gram, punctuation and pragmatic features	Kunneman et al. (2015)
Bag of word and N-gram	Ling and Klinger (2016)
Lexical, subjectivity, N-gram, word2vec	Schifanella et al. (2016)
N-gram and sentiment based features	Bouazizi and Ohtsuki (2016)
BOW, POS and sentiment feature	Ghosh and Veale (2016)
N-gram, Bag of word	Amir et al. (2016)
Contextual features	Zhang et al. (2016)
Behavioral features (Likes and dislikes)	Bharti et al. (2016)
sentiment and emotion-based features	Sulis et al. (2016)
Content word, function word, N-gram and parts of speech	Mukherjee and Bala (2017b)
Sentiment-based feature	Manohar and Kulkarni (2017)
Punctuation mark, Dots, positive words and bracket	Al-Ghadhban et al. (2017)
Sentiment polarity feature	Ranjan et al. (2017)
Function words, content words and parts of speech N-gram	Mukherjee and Bala (2017a)
N-gram, sentiment, topic	Dharwal et al. (2017)
Interjections and intensifiers	Bharti et al. (2017)
Hyperbolic, question mark and intensifiers	Abulaish and Kamal (2018)
lexical, pragmatic, hyperbole, quotations and punctuation marks	Samonte et al. (2018)
N-gram, sentiment and emoticon	Sreelakshmi and Rafeeque (2018)
Pragmatic, N-gram and sentiment features	Manjusha and Raseek (2018)
N-gram	Kumar and Harish (2018)
Punctuation mark, capital letter and ‘or’ conjunction	Kumar et al. (2019)
Pragmatic features, Malay prosodic, syntactic feature, POS tagger	Suhaimin et al. (2019)

improved the accuracy of the prediction. The summary of the features extraction on the selected studies is shown in Table 8.

5.3.2 Review of feature representation techniques

In addition to the feature extraction techniques, the study revealed that the feature representation techniques mostly used to convert the extracted feature into numerals is term frequency (TF), which is used to determine the frequency and occurrence of sarcasm in the extracted features. For instance, the contextual features extracted from the target author's historical tweets in a study by Suhaimin et al. (2019) were represented with TF and IDF. In that regard, the feature values of TF-iDF were used to sort the history tweets in order to choose the constant number of contextual tweets word (feature), having the greatest values of TF-iDF. In another study, Suhaimin et al. (2019), on sarcasm detection and sentiment analysis classification, the three NLP categories of features (pragmatic, syntactic, and prosodic), proposed by Suhaimin et al. (2018), were adopted due to the demonstration of its improvement in sarcasm detection. Thus, the extracted features were represented using term frequency-inverse document frequency (TF-iDF) and binary representation (BR). Out of the 40 selected studies, 12 studies used TF, 8 studies used BR and 20 studies did not report any feature representation technique that was used.

5.3.3 Review of feature selection techniques for sarcasm identification

In feature selection, certain criteria are followed to discover suitable feature sets (Guyon and Elisseeff 2003) and it is broadly employed in sarcasm detection. Notwithstanding, only a few studies in the selected studies on sarcasm identification utilized the feature selection technique to investigate the outcome of the different subgroups on the classification accuracy. The feature selection techniques that were used on the selected studies are Chi square (χ^2), information gain (IG) and mutual information (MI), which are briefly explained below.

Chi square (χ^2) Chi square is a statistical test used for measuring the absence of the independence that exists between a particular class (c) and term of features (f) (Kumar and Harish 2018).

Information gain (IG) Information gain is a feature selection technique that is used to determine the information gain by knowing the value of the attribute within a feature vector (Yang and Pedersen 1997).

Mutual information (MI) It is a statistical measure that is commonly used to model two random variables (word association and related application) that are mutually dependent (Yang and Pedersen 1997).

For instance, Kumar and Harish (2018) employed Chi square (χ^2), mutual information (MI), and information gain (IG) as conventional feature selection techniques to select the discriminative features for sarcasm classification. The researcher tested their presence and the experimental finding shows that the use of these feature section techniques brought about the reduction of the high dimensional feature space and also increased the classifiers classification accuracy. For example, SVM and RF classifiers yielded a maximum accuracy when MI and IG selection scheme were applied in classification. In a related study Muresan et al. (2016), the N-gram lexical features were extracted using linguistic inquiry and word count (LIWC) and WordNet-Affect dictionary (Strapparava and Valitutti 2004; Pennebaker et al. 2015). Furthermore, pragmatic

features such as emoticon and punctuation were extracted. However, the discriminative features were selected in these features by employing the Chi square (χ^2) selection scheme before modelling. The review showed that five (5) out of the 40 selected studies used Chi square to select discriminative features, three (3) studies used information gain, one study used Chi square, information gain and mutual information (MI), 31 studies, however, did not report the use of any feature selection scheme to select the important feature from the extracted one. The summary of the feature representation techniques is shown in Table 7, while the feature selection scheme utilized in the analyzed studies is shown in Table 8.

5.4 Review of classification techniques for sarcasm identification

Various classification algorithms according to our findings have been used for sarcasm identification in the social media. The review summary of the classification algorithms used in the selected studies is depicted in Table 9, which shows that one or more classifiers have been utilized by each study. In addition, some studies utilized multiple classifiers in order to compare the performance of each classifier with the proposed method. It is obvious from Table 9 that some studies employed only one learning algorithm for classification. Moreover, different researchers on sarcasm identification used different datasets. Thus, the comparison of different classifiers performance in classification in such an instance becomes difficult. For instance, a few recent studies Liebrecht et al. (2013) and Kunneman et al. (2015) employed only balanced winnow classifiers for sarcasm identification. In these studies, a balanced winnow allocates scores to each class label and good performance was obtained when area under curve (AUC) metrics were used, which showed its confidence in such a label. In another study, random forest (RF), support vector machine (SVM), K-nearest neighbour (K-NN) and maximum entropy (ME) were used to classify sarcasm on tweets datasets using pattern related features. The performance classifier result showed that RF outperformed SVM, K-NN and ME by attaining an accuracy of 81.3% F-measure. Ling and Klinger (2016) in their study on the 'Comparative analysis classification of differences between irony and sarcasm', compared the performance of the DT, ME and SVM classifiers. The empirical analysis showed that the ME model performed better than the decision tree and SVM classifiers. Sulis et al. (2016), investigated the classifier performance of Naïve Bayes (NB), DT, RF, LR and SVM in modelling the differences among the three figurative messages (#sarcasm, #Not and #Irony) on twitter. Among these classifiers, the highest result of f-measure was obtained by applying RF classifier in distinguishing #Irony versus #Not. However, when similar datasets used in (Barbieri et al. 2014) were employed for the #Irony versus #Sarcasm classification experiment, the performance result showed an improvement of F-measure from 0.62 to 0.70. Moreover, Abulaish and Kamal (2018) compared the performance of NB, DT and Bagging (ensemble) classifier to classify hyperbolic and self-deprecating features for sarcasm identification in the tweets datasets (balanced and unbalanced). They reported the performance result of the experiment in the form of precision, f-measure and recall in applying all the three classifiers, that the DT attained highest values in f-measure and recall while the best precision value was achieved by the bagging classifier in both datasets. It is obvious from Table 9 that support vector machine (SVM) and Naïve Bayes (NB) are the most used classifiers for sarcasm identification in the social platform. Among the 40 selected studies, 22 used the SVM classifier and 14 used NB (Fig. 7).

Table 7 Feature representation techniques used on the selected studies

Feature representation technique	Count	References
BR	9	Riloff et al. (2013), Liu et al. (2014), Ghosh et al. (2015), Khattri et al. (2015), Amir et al. (2016), Schifanella et al. (2016), Sulis et al. (2016), Mukherjee and Bala (2017a) and Sreelakshmi and Rafeeque (2018)
TF	5	Davidov et al. (2010), González-Ibánez et al. (2011), Liebrecht et al. (2013), Kumar and Harish (2018) and Manjusha and Raseek (2018)
TF-iDF	4	Ptáček et al. (2014), Dharwal et al. (2017), Samonte et al. (2018) and Suhaimin et al. (2019)
BR and TF	1	Barbieri et al. (2014)
TF and TF-iDF	2	Zhang et al. (2016) and Ranjan et al. (2017)
The Feature representation technique that was used was not mentioned	19	Lunando and Purwarianti (2013), Altrabsheh et al. (2015), Bharti et al. (2015), Bouazizi and Ohtsuki (2015b), Fersini et al. (2015), Kunneman et al. (2015), Rajadesingan et al. (2015a), Wang et al. (2015), Bharti et al. (2016), Bouazizi and Ohtsuki (2016), Ghosh and Veale (2016), Ling and Klinger (2016), Al-Ghadhban et al. (2017), Bharti et al. (2017), Manohar and Kulkarni (2017), Mukherjee and Bala (2017b), Abulaish and Kamal (2018) and Kumar et al. (2019)

Table 8 Feature selection techniques used on the selected studies

Feature selection technique	Count	Reference
Chi square	5	González-Ibánez et al. (2011), Liebrecht et al. (2013), Dharwal et al. (2017), Manjusha and Raseek (2018) and Sreelakshmi and Rafeeque (2018)
Information gain	3	Barbieri et al. (2014), Liu et al. (2014) and Sulis et al. (2016)
Chi square, Information gain and Mutual information	1	Kumar and Harish (2018)
The Feature selection technique that were used was not mentioned	31	Davidov et al. (2010), Lunando and Purwarianti (2013), Riloff et al. (2013), Ptáček et al. (2014), Altrabsheh et al. (2015), Bharti et al. (2015), Bouazizi and Ohtsuki (2015a, b), Fersini et al. (2015), Ghosh et al. (2015), Khattri et al. (2015), Kunneman et al. (2015), Rajadesingan et al. (2015a), Wang et al. (2015), Amir et al. (2016), Bharti et al. (2016), Bouazizi and Ohtsuki (2016), Ghosh and Veale (2016), Ling and Klinger (2016), Schifanella et al. (2016), Zhang et al. (2016), Al-Ghadhban et al. (2017), Bharti et al. (2017), Manohar and Kulkarni (2017), Mukherjee and Bala (2017a, b), Ranjan et al. (2017), Abulaish and Kamal (2018), Samonte et al. (2018), Kumar et al. (2019) and Suhaimin et al. (2019)

5.5 Review of performance measure

The performance evaluation of sarcasm classification can be measured using various performance metrics such as accuracy (ACC), recall (REC), F-measure (F-M), precision (PR), the Area under curve (AUC) and kappa statistics (KS). The values of false positive (FP), false negative (FN), true positive (TP), and true negative (TN), which are the contents of the confusion matrix can be used for computation of these metrics. The detail description and the computation of these measures are given in Sect. 3.7. However, the choice of selecting the performance metrics depends on the goal for which sarcasm is being identified. Although the review indicated precision, accuracy, recall, and F-measure as the mostly employed performance metrics, these metrics may be inadequate to correctly evaluate the classifier's performance correctly. This is because of the class imbalance in various datasets found in most selected studies. In such a situation, AUC would be the best option due to its suitability in evaluating the classification performance related to an individual class (Provost and Fawcett 1997; Provost et al. 1998). For instance, Samonte et al. (2018) collected two sets of tweets dataset (English and Filipino) on a range of domains such as social media, politics, weather, government and transportation to build a model for sarcasm identification in a multilingual platform. In the study, only accuracy metrics were employed by the author to measure the performance of the classification. The English datasets comprised 1101 sarcastic and 13,998 non-sarcastic, whereas Filipino datasets consisted of 894 sarcastic and 14,229 non-sarcastic. Here, the two sets of data are naturally imbalance and in such a case, there may be biases in using the only accuracy as performance metrics. Thus, the right measure to accurately determine the performance of the algorithm for sarcasm identification is AUC. In another study, Liu et al. (2014) employed two corpora to classify English and Chinese sarcasm features. The first corpus consists of Twitter, Amazon product review and News article datasets. Among this corpus, the Twitter dataset comprised 3200 sarcastic and 36,800 non-sarcastic, Amazon product (471 sarcastic

Table 9 Classification algorithm used on the selected studies

Studies	SVM	NB	RF	ME	DT	LR	KNN	ANN	FC	RB	BAGGING	AB	BW
Davidov et al. (2010)							✓						
González-Ibánez et al. (2011)	✓					✓							
Lunando and Purwarianti (2013)	✓	✓		✓									
Liebrecht et al. (2013)													✓
Riloff et al. (2013)	✓												
Ptáček et al. (2014)					✓								
Barbieri et al. (2014)	✓			✓									
Liu et al. (2014)	✓	✓		✓									
Bouazizi and Ohtsuki (2015b)			✓										
Fersini et al. (2015)	✓	✓			✓								
Khattri et al. (2015)										✓			
Ghosh et al. (2015)	✓												
Rajadesingan et al. (2015b)	✓				✓	✓							
Bharti et al. (2015)													
Bouazizi and Ohtsuki (2015a)	✓	✓		✓									
Altrabsheh et al. (2015)		✓	✓	✓						✓			
Wang et al. (2015)	✓												
Kunneman et al. (2015)													✓
Ling and Klinger (2016)	✓			✓	✓								
Schifanella et al. (2016)	✓												
Bouazizi and Ohtsuki (2016)	✓		✓	✓			✓						
Ghosh and Veale (2016)	✓							✓					
Amir et al. (2016)								✓					
Zhang et al. (2016)								✓					
Bharti et al. (2016)													

Table 9 (continued)

Studies	SVM	NB	RF	ME	DT	LR	KNN	ANN	FC	RB	BAGGING	AB	BW
Sulis et al. (2016)	✓	✓	✓		✓	✓							
Mukherjee and Bala (2017b)		✓							✓				
Manohar and Kulkarni (2017)													
Al-Ghadhban et al. (2017)		✓											
Ranjan et al. (2017)	✓	✓											
Mukherjee and Bala (2017a)		✓		✓									
Dharwal et al. (2017)	✓					✓							
Bharti et al. (2017)	✓	✓	✓		✓							✓	
Abulaish and Kamal (2018)		✓			✓						✓		
Samonte et al. (2018)	✓	✓		✓									
Sreelakshmi and Rafeeque (2018)	✓				✓								
Manjusha and Raseek (2018)	✓	✓					✓	✓					
Kumar and Harish (2018)			✓										
Kumar et al. (2019)								✓					
Suhaimin et al. (2019)	✓												
Total:	22	14	6	9	8	4	3	5	1	2	1	1	2

RF random forest; *ME* maximum entropy; *SVM* support vector machine; *DT* decision tree; *LR* logistic regression; *NB* Naive Bayes; *ANN* artificial neural network; *KNN* K-nearest neighbour; *FC* fuzzy clustering; *RB* rule base; *AB* Adaboost; *BW* balanced winnow

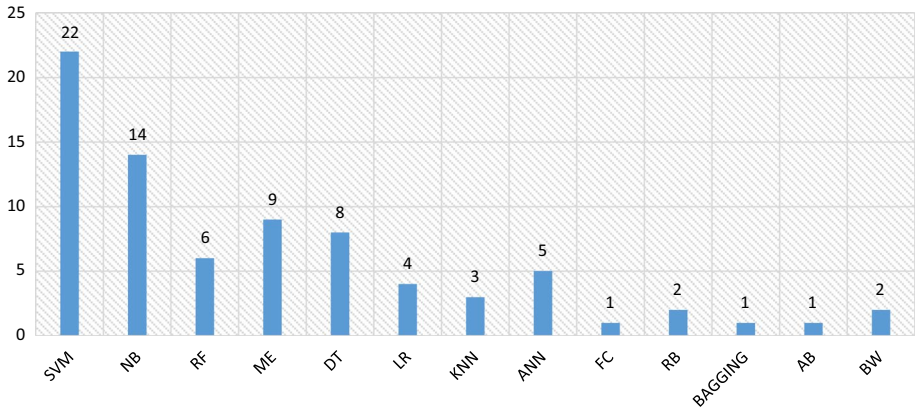

Fig. 7 Frequency of the classification techniques used in the selected studies

and 5020 non-sarcastic), News article (223 sarcastic and 4000 non-sarcastic). However, the second corpus consist of three Chinese topic comments crawled from Tencent Weibo (359 sarcastic and 5128 non-sarcastic), Sina Weibo (238 sarcastic and 3621 non-sarcastic), and Netease BBC (546 sarcastic and 9810 non-sarcastic). It is obvious that all the class distributions of the corpus used in the classification experiment are highly imbalanced. Thus, area under the curve (AUC) was employed by the authors to accurately measure the performance of the classification models. This is because; AUC has a strong resistance to the skewness in datasets compared to the F-score, when employing TPR instead of precision. The summary of the performance measure used in the selected studies is shown in Table 10.

5.6 Discussion

The extensive review of the academic articles on sarcasm identification classification published between 2008 and 2019 has been carried out in this study. The review concentrated on the aspect of dataset usage, the pre-processing techniques, the feature engineering techniques, the classification algorithm and the performance measures used in the selected studies. It was discovered in the study that sarcasm detection has been applied in many application domains such as product review, sentiment analysis, spam email filtering, and dialogue in human–computer interaction, etc.

The first review question: "Are there annotated sarcastic datasets publically available in the area of sarcasm identification using text classification methods?", provide insight on various publicly available datasets for sarcasm identification. The review findings show that the datasets for sarcasm identification are obtained by researchers due to the fact that there is no standard publicly available datasets on sarcasm identification except the Amazon product review datasets, which are only available only on request. Many studies have collected their datasets on a microblogging sites such as Twitter. The distinctiveness properties of Twitter have made it to be the mostly utilized in comparison with other type of datasets. Some of the reasons for employing twitter include the generation of the large volume of the tweets in a short period of time, as Twitter data consists of different characteristics that can be categorized when crawling the data such as the domain type, trending (past and current trends), politics, gender, age factors, and geographical location. In addition, the

Table 10 The frequency of performance metrics in the selected studies

Studies	ACC	PR	REC	F-M	AUC	KS
Davidov et al. (2010)	✓	✓	✓	✓		
González-Ibánez et al. (2011)	✓					
Lunando and Purwarianti (2013)	✓					
Liebrecht et al. (2013)	✓		✓		✓	
Riloff et al. (2013)		✓	✓	✓		
Ptáček et al. (2014)		✓	✓	✓		
Barbieri et al. (2014)				✓		
Liu et al. (2014)					✓	
Bouazizi and Ohtsuki (2015b)	✓	✓	✓			
Fersini et al. (2015)	✓	✓	✓	✓		
Khattri et al. (2015)		✓	✓	✓		
Ghosh et al. (2015)		✓	✓	✓		
Rajadesingan et al. (2015b)	✓				✓	
Bharti et al. (2015)		✓	✓	✓		
Bouazizi and Ohtsuki (2015a)	✓	✓	✓	✓		
Altrabsheh et al. (2015)	✓	✓	✓	✓	✓	
Wang et al. (2015)	✓					
Kunneman et al. (2015)		✓			✓	
Ling and Klinger (2016)	✓					
Schifanella et al. (2016)	✓					
Bouazizi and Ohtsuki (2016)	✓	✓	✓	✓		
Ghosh and Veale (2016)		✓	✓	✓		
Amir et al. (2016)	✓					
Zhang et al. (2016)		✓	✓	✓		
Bharti et al. (2016)		✓	✓	✓		
Sulis et al. (2016)				✓		
Mukherjee and Bala (2017b)	✓	✓	✓	✓		
Manohar and Kulkarni (2017)	✓					
Al-Ghadhban et al. (2017)	✓	✓	✓	✓		
Ranjan et al. (2017)	✓	✓	✓	✓		
Mukherjee and Bala (2017a)	✓	✓	✓	✓		
Dharwal et al. (2017)				✓		
Bharti et al. (2017)		✓	✓	✓		
Abulaish and Kamal (2018)		✓	✓	✓		
Samonte et al. (2018)	✓	✓	✓	✓		✓
Sreelakshmi and Rafeeque (2018)	✓	✓	✓	✓		
Manjusha and Raseek (2018)		✓	✓	✓		
Kumar and Harish (2018)	✓	✓	✓	✓		
Kumar et al. (2019)	✓	✓	✓	✓		
Suhaimin et al. (2019)				✓		

ACC accuracy; *PR* precision; *REC* recall; *F-M* F-measure; *AUC* area under the curve; *KS* kappa statistics

use of #hashtag and keyword for streaming on Twitter is another property that is of great interests to researchers in using the Twitter domain. A Twitter hashtag is a string preceded by the hash symbol, which can be viewed as a topic marker or the key context expression of the tweet. Thus, users that discuss similar topics make use of the hashtag (Tsur and Rappoport 2012). One of the issues observed with the datasets deployed in the studies is due to the imbalanced nature of the datasets. In such studies, there is an unequal distribution of class instances, which can result in the bias of the classification accuracy. Conversely, the review showed that there is no publicly available annotated datasets in this research domain. Therefore, it is necessary to have a standard public datasets for the classification experiment on sarcasm identification and to employ the suitable performance metrics such as AUC for the evaluation of classification performance when an imbalance datasets are used.

Furthermore, various pre-processing techniques have been employed to process the extracted data in order to cleanse the data from the unwanted item that will not contribute to the classification performance. Nevertheless, only few studies (Riloff et al. 2013; Al-Ghadhban et al. 2017; Bharti et al. 2017; Mukherjee and Bala 2017b; Ranjan et al. 2017; Manjusha and Raseek 2018; Samonte et al. 2018) experimented the existence and non-existence of the stop word and reported that the removal of stop word attained a better accuracy in classification than their presence. Also, some of the selected studies indicated that the application of word tokenization with basic pre-processing task achieved better performance in classification (Riloff et al. 2013; Barbieri et al. 2014; Ghosh et al. 2015; Khattri et al. 2015; Ranjan et al. 2017; Samonte et al. 2018). In some selected studies, stemming (Riloff et al. 2013; Al-Ghadhban et al. 2017; Dharwal et al. 2017; Samonte et al. 2018) and lemmatization (Bouazizi and Ohtsuki 2016; Manohar and Kulkarni 2017) have also been applied and the effectiveness of the techniques has been demonstrated. Besides, researchers have also demonstrated the text normalization technique. In such a technique, the data were scaled to a common unit using a regular expression. The research finding shows that text normalization helps in the improvement of the classification performance and therefore eliminates the dimensionality problem (Patro and Sahu 2015). As such, there is a need for empirical evaluations and comparison of some of the pre-processing techniques on the collected data for sarcasm identification so as to ensure better classification performance.

The review answered the next four research questions (*Research Question* 2, *Research Question* 3, *Research Question* 4 and *Research Question* 5 as outlined in Sect. 1. The research questions seek to answer various feature engineering techniques (consisting of feature extraction, feature selection, and feature representation) employed in the selected studies. Based on the findings from the review, most researchers employed the feature extraction techniques that consist of N-gram, BoW, Word2vec, and PoS tagging technique to extract discriminative features from the collected sarcastic datasets before the classification stage. However, as revealed in Table 6, most studies utilized N-gram extraction technique for sarcasm identification due to its simplicity and scalability properties. Thus, content-based linguistic features such as unigram, bigram, trigram, among others, were most useful features in the selected studies for sarcasm identification. In sarcasm identification, it is not encouraged to rely only on the content-based features extraction for classification. This is because of the limited accuracy that may occur in the classification performance due to the limitations inherent in those features. One of the issues with the content-based feature is disregarding of word order and grammar even though the word frequency is retained. Secondly, these features do not account for the word-level synonyms and polysemy when used for sarcasm identification. In order to avoid these limitations, a combination of other

features together with the content-based feature is necessary to enhance the classification accuracy. In addition to the feature extraction, several studies used binary representation (BR) to find the occurrence of sarcasm on the extracted feature and term frequency (TF) representation scheme to identify the frequency occurrence of the sarcastic features in the extracted feature. A study by Barbieri et al. (2014) represented sarcastic features using TF and BR and obtained a promising result. Therefore, TF and BR are mostly employed feature representation techniques in the selected studies and are thus, recommended for sarcastic feature representation due to the promising results obtained on the studies that have used them. It should also be noted that not all the available features might be useful in realizing improved classification performance accuracy since indiscriminative features may lead to model over-fitting (Forman 2003). Hence, suitable feature selection scheme is required in order to find the useful features that can enhance the classification accuracy, lower the computation time and decrease the noise in the construction of the classification model (Hall and Smith 1998). Consequently, the review on the selected study indicated that Chi square (χ^2), information gain (IG) and mutual information (MI) feature selection schemes were mostly employed for the selection of relevant features.

In answering the Research Question 6: "Which of the text classification algorithms produces better accuracy and why?", the review discovered that various classification algorithms have been employed in the selected studies for identification of sarcasm in social media platforms. However, the result of the analysis in the studied datasets with the proposed features in their corresponding studies showed that SVM produced the best performance results (González-Ibánez et al. 2011; Riloff et al. 2013; Ghosh et al. 2015; Schifanella et al. 2016). For instance, González-Ibánez et al. (2011) in their study tested the evaluation of SVM and logistic regression classifier for classification in order to distinguish sarcasm from the positive and negative sentiment in the Twitter message after using the Chi squared feature selection scheme to select the most discriminant feature; and it was reported that the accuracy outperformed the LR model. Recently, Riloff et al. (2013) carried out the comparison of the SVM classifier and rule-based approach in the detection of sarcasm and produced a better result than using only the ruled based approach. Interestingly, the sparse nature of the SVM model has made it suitable for text classification. Report on several studies also indicated that NB algorithm produced enhanced classification results in sarcasm identification. Furthermore, only a few studies among the selected studies applied the KNN algorithm for sarcasm detection and the experimental results in those studies showed vacillating results. Nonetheless, the analysis of different results on the selected studies showed that SVM produced better performance in sarcasm classification followed by NB, and KNN classification algorithms as they also provided optimum performance in the selected studies. It should be noted that four (4) studies (Amir et al. 2016; Ghosh and Veale 2016; Zhang et al. 2016; Manjusha and Raseek 2018) out of the 40 selected studies used deep learning approach for sarcasm classification and compared the result of the deep learning with the traditional machine learning approach such as LR, SVM and RF. The results of the experiments showed that deep learning outperformed traditional machine learning. For example, a novel convolutional network-based approach was presented by Amir et al. (2016), the study learnt the user-specific context and reported a 2% improvement in performance. In addition, Ghosh and Veale (2016) combined convolutional neural network (CNN), deep neural network (DNN) and long short term memory (LSTM) in their classification approach, thus, resulting to an improvement shown by their deep learning architecture, as when compared with the recursive support vector machine model. The main advantage of deep learning is that feature is engineered and learned automatically through a general learning process, unlike the shallow learning that depends on

a human for feature engineering. Thus, the deep learning approach is very helpful in sarcasm detection classification by solving the problem of data dimensionality, which usually occurs when features are humanly engineered.

From the Research Question 7: "Which performance measures are most widely used to measure the performance of the classifiers in sarcasm classification?", the analysis of the selected studies indicated that precision, accuracy, recall, and F-measure were the mostly employed performance metrics yet, these metrics may be inadequate to correctly evaluate the classifier performance. This is because of the class imbalance that is mostly found in various datasets in the selected studies. In such a situation, AUC would be the best option due to its suitability in evaluating the classification performance related to an individual class (Provost and Fawcett 1997; Provost et al. 1998). Besides, AUC has a feature of strong resistance to the skewness in datasets by using TPR when compared with F-Measure.

Based on the review, only one study (Ptáček et al. 2014) out of the 40 selected studies provided a detail error analysis for misclassification. For instance, in the study for sarcasm detection on English and Czech tweets, an imbalanced distribution performance was carried out. In their experiment, an English corpus consisting of 100,000 tweets was sampled to obtain similar distribution on Czech corpus consisting of 325 sarcastic and 6675 non-sarcastic tweets. Thus, the combination of various features yields F-measure of 0.734 ± 0.01 on the Maximum Entropy classifier and 0.729 ± 0.01 on SVM which shows the drop in the performance. This is an indication that the amount of training data plays a vital role in classification performance (0.92 approximation on English corpus versus 0.73 approximation on Czech corpus). Hence, wrong classification may lead to poor performance. To this end, research questions 1 to 7 of this study have been answered while research question 8 is answered in Sect. 6 below.

6 Research challenge and future directions

This review has identified several research issues inherent in the previous researches in sarcasm identification using text classification approaches. The highlighted research gaps need considerable research efforts to create an efficient classification model in the domain of sarcasm identification. These research challenges require further research in order to solve them. These challenges and open research directions are discussed below:

1. *Datasets* One of the major problems in sarcasm identification domain is lack of standard dataset. There is no standard publicly available dataset for sarcasm identification; this has made most researchers to create privately owned datasets. Consequently, this situation has resulted in the biases of the data since both the training and testing sets are created by the researchers and there is no existing standard data that can be used for comparison with the proposed technique to evaluate the unbiased in terms of the performances. There is also an imbalance in the class distribution of the datasets which make the number of sarcastic text data and non-sarcastic not to correspond to the same size. This calls for the creation of standard datasets, which will solve the problem of biases in the data. A technique also needs to be proposed in order to balance the datasets before classification experiment and to apply performance metrics such as AUC, which is suitable for the evaluation of the performance of the classifier in the imbalance datasets.

2. *Tweets typo* Twitter data is the most widely used domain for sarcasm detection according to our review. Misspelling of words has become a common mistake in microblog while composing a message. Humans, without any effort, can easily correct such mistakes manually but it is very difficult for machine learning to detect and correct such misspelt words. However, such words can correspond to a specific dictionary that has been removed during the pre-processing stage. Thus, it can drastically influence the sentence polarity. Not only that, machine learning could ignore such wrongly spelt words and replace them with closely related ones. Notwithstanding, such errors are very common in sarcasm detection. Thus, attention should be paid in finding a technique that could detect and correct such wrongly spelt words.
3. *The exploitation of new features* The review shows that most of the existing studies made use of the content-based linguistic-based features in the classification phase for sarcasm identification on social media platform. However, only a few studies (Bharti et al. 2016; Zhang et al. 2016) took advantage of the behavioural and contextual features to identify sarcasm. In those studies, promising accuracies were obtained compared with the content-based features. One of the studies Schifanella et al. (2016), out of the 40 selected studies also made use of the visual semantics feature (VSF), in which the sarcasm can only be understood through the semantics in the image and was able to attain a higher accuracy when combined with N-gram with the SVM classifier. Therefore, it is important for future research to explore various novel features such as behavioural, contextual and visual features for sarcasm identification.
4. *Application of deep learning methods* Most researchers in the field of data mining domain are now shifting from the traditional machine learning to Deep learning methods due to the cumbersomeness inherent in the pre-classification phase especially the feature extraction phase in the traditional machine learning approaches for sarcasm identification. The deep learning approach is required in order to overcome such issues, as the features are not engineered by human intervention. Only four (4) studies (Amir et al. 2016; Ghosh and Veale 2016; Zhang et al. 2016; Manjusha and Raseek 2018) out of the 40 selected studies made use of deep learning approach. The classification accuracy of the sarcasm detection can be enhanced by applying different deep learning techniques for effective feature extractions such as word to vector (word2vec) conversion, n-gram and bag-of-words. Some of the deep learning classification algorithms such as recurrent neural networks (RNN) and convolutional neural network (CNN), have reported good performance when applied in sarcasm identification. Deep learning has also enhanced the performance accuracy in many texts and web mining classification (Dumais and Chen 2000). As such, future research can shift attention to the application of deep learning methods.
5. *Intense use of emoji and emoticon* People have been familiar with the use of emotion symbols like emoji and emoticon in the social media to display their state of mind especially in microblog that has restrictions on the number of characters per chat. Ambiguity is likely to occur among the users with regards to the specific meaning of emoji. Thus, it has the ability to change the overall sentiment of the sentence as the emoji features are not incorporated into the current system. To this end, future researchers should take note of how to investigate and incorporate these features.
6. *Multilingual-based approach* Majority of the existing works on sarcasm identification utilized only English language datasets. However, most people usually express their emotions better in their native languages than in English. Thus, mining such opinions becomes problematic because many people do not have interest in such research; that is why most existing works on sarcasm classification paid more attention to textual data

expressed only in English language. However, only a few studies worked on the other languages apart from English. For instance, (Samonte et al. 2018) in their study worked on the sentence level sarcasm detection in English qand Filipino tweets. Classification performances were compared in both languages and the result showed that maximum entropy (ME) model obtained a better accuracy of 88.506% for training and 91.994% after validation when applied on Filipino datasets compared with English datasets that produced an accuracy of 79.91% for training and 78.75% for testing. As such, further research that will focus on feature extraction on other languages and modification of classifiers is urgently required so that it can be applicable in sarcasm identification written in other languages.

7. *Clustering-based approach* Clustering-based approach deploy an unsupervised learning approach (Yang 1993) that is mostly applicable in pattern recognition but this is still an infant in the domain of sarcasm identification. Most researchers in the selected studies implemented a supervised learning approach to build a classification model and obtained a good result despite the limitations inherent in such approaches. One of the key issues in supervised learning is the labelling of the datasets in order to construct the training sets. Such tasks require linguistic experts and they are time-consuming. For instance, in a study conducted by Samonte et al. (2018) for detection of sarcasm in English and Filipino at the sentence level, six (6) experts in the linguistics were engaged to manually label 30,231 tweets (that consists of 15,099 English and 15,132 Filipino) as sarcastic or non-sarcastic. Thus, a tremendous amount of time is required in the preparation, and disagreement could arise in a situation where more than an expert is engaged for annotation. So, further research in this domain can focus more on the unsupervised approach (clustering) for modelling sarcasm identification in order to get rid of such labelling exertion.

7 Conclusion

The study presents a comprehensive review of classification techniques for sarcasm identification on the social media platform. The comprehensive review covered articles on sarcasm detection published between 2008 and 2019. The study selected 40 primary studies from 7 different academic databases and critically reviewed the areas of datasets usage, pre-processing techniques, feature engineering techniques (consisting of feature extraction, representation, and selection), the classification approach and the performance metrics. The study showed that there are no standard and publicly available datasets for sarcasm identification in social microblogs such as Twitter in such a way that researchers are required to crawl their own datasets. Content-based features were mostly used features whereas N-gram and POS tagger were the mostly used feature extraction techniques due to their simplicity in usage. (BR) and TF were the most used feature representation schemes in the selected studies. BR technique is very effective in sentiment feature representation, as the occurrence of the sarcasm is checked on the textual data. For example, sentiment 1 is used to indicate the presence of sarcasm in the sentence whereas sentiment 0 indicates the absence of sarcasm in the sentence. TF was also used to check the frequency of occurrence of the feature in the training sets; this has the potential of increasing the likelihood occurrence of feature in the test set. In order to eliminate the non-discriminative features, various studies applied feature selection schemes such as Chi squared and Information gain. In the classification phase, the majority of the studies applied supervised

machine learning algorithms such as SVM, NB, RF, ME and DT. The review showed that the SVM algorithm is mostly used, followed by NB, RF, and ME. This is so because it obtained better result compared to other classifiers. Only a few studies used rule-based and NLP approaches. In recent studies, a deep learning approach has gained ground in sarcasm identification owing to the fact that learning and feature engineering is done automatically without human intervention. Performance metrics such as precision, recall, accuracy, and F-measure were used as a performance measure to measure the performance of the classification algorithm and it was found that accuracy was mostly used in the selected studies. Relying only on the accuracy for performance measure will not produce a better result in a situation where imbalance datasets are used. Hence, AUC is a more suitable metrics for performance measure where there are datasets imbalances. A comprehensive investigation of characteristics, types, strengths, and weaknesses of datasets for sarcasm identification in the social media textual data was carried out. In addition, outline taxonomy, various features representation and extraction for efficient algorithm development are presented. The survey also critically analyzed various data preparation (pre-processing) techniques and recent classification algorithms for sarcasm identification. Finally, in order to set the pace for development of the new ground, the study identifies recent research challenges and proposes open research direction to tackle issues in sarcasm identification domain. This comprehensive review of sarcasm identification systems would provide invaluable insight into the research domain and researchers are to further improve sarcasm identification system using textual data.

Appendix: The following abbreviations and their full form were used in this paper

Abbreviations	Definitions	Abbreviations	Definitions
AB	Adaboost	ME	Maximum entropy
ACC	Accuracy	MI	Mutual information
ANN	Artificial neural networks	NB	Naïve Bayes
API	Application protocol interface	NLP	Natural language processing
AUC	Area under the curve	POS	Part of speech tagging
BoW	Bag of words	PRE	Precision
BR	Binary representation	RB	Rule base
CNN	Convolutional neural network	REC	Recall
BW	Balanced winnow	RF	Random forest
CUE-CNN	Convolutional user embedding convolutional neural network	RNN	Recurrent neural network
DNN	Deep neural network	SLR	Systematic literature review
DT	Decision tree	SMO	Sequential minimal optimization
FC	Fuzzy clustering	SVM	Support vector machine
F-M	F-measure	TF	Term frequency
FN	False negative	TFIDF	Term frequency with inverse document frequency
FP	False positive	TN	True negative
IG	Information gain	TP	True positive

Abbreviations	Definitions	Abbreviations	Definitions
KS	Kappa statistics	TPR	True positive rate
k-NN	k-nearest neighbours	URL	Universal resource locator
LSTM	Long short term memory	VSF	Visual semantic feature
LR	Logistic regression		

References

Abercrombie G, Hovy D (2016) Putting sarcasm detection into context: the effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. Paper presented at the Proceedings of the ACL 2016 Student Research Workshop

Abulaish M, Kamal A (2018) Self-deprecating sarcasm detection: an amalgamation of rule-based and machine learning approach. Paper presented at the 2018 IEEE/WIC/ACM international conference on web intelligence (WI)

Al-Ghadhban, D., Alnkhilan, E., Tatwany, L., & Alrazgan, M. (2017). Arabic sarcasm detection in Twitter. Paper presented at the 2017 International Conference on Engineering & MIS (ICEMIS)

Altrabsheh N, Cocea M, Fallahkhair S (2015) Detecting sarcasm from students' feedback in Twitter. In: Design for teaching and learning in a networked world. Springer, Cham, pp 551–555

Amir S, Wallace BC, Lyu H, Silva PCMJ (2016). Modelling context with user embeddings for sarcasm detection in social media. arXiv preprint arXiv:1607.00976

Barbieri F, Saggion H, Ronzano F (2014). Modelling sarcasm in twitter, a novel approach. Paper presented at the proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis

Bharti SK, Babu KS, Jena SK (2015) Parsing-based sarcasm sentiment recognition in Twitter data. Paper presented at the proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015—ASONAM '15

Bharti S, Vachha B, Pradhan R, Babu K, Jena S (2016) Sarcastic sentiment detection in tweets streamed in real time: a big data approach. Digit Commun Netw 2(3):108–121

Bharti SK, Naidu R, Babu KS (2017) Hyperbolic feature-based sarcasm detection in tweets: a machine learning approach. Paper presented at the 2017 14th IEEE india council international conference (INDICON)

Bouazizi M, Ohtsuki T (2015a) Opinion mining in Twitter: how to make use of sarcasm to enhance sentiment analysis. Paper presented at the 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)

Bouazizi M, Ohtsuki T (2015b) Sarcasm detection in Twitter: "all your products are incredibly amazing!!!"—are they really? Paper presented at the 2015 IEEE global communications conference (GLOBECOM)

Bouazizi M, Ohtsuki TO (2016) A pattern-based approach for sarcasm detection on twitter. IEEE Access 4:5477–5488

Burfoot C, Baldwin T (2009) Automatic satire detection: are you having a laugh? Paper presented at the proceedings of the ACL-IJCNLP 2009 conference short papers

Cotelo JM, Cruz FL, Troyano JA, Ortega FJ (2015) A modular approach for lexical normalization applied to Spanish tweets. Expert Syst Appl 42(10):4743–4754

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

da Silva NFF, Hruschka ER, Hruschka ER (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66:170–179. https://doi.org/10.1016/j.dss.2014.07.003

Dai Q-Y, Zhang C-P, Wu H (2016) Research of decision tree classification algorithm in data mining. Int J Database Theory Appl 9(5):1–8

Davidov D, Tsur O, Rappoport A (2010) Semi-supervised recognition of sarcastic sentences in twitter and amazon. Paper presented at the Proceedings of the fourteenth conference on computational natural language learning

Debole F, Sebastiani F (2004) Supervised term weighting for automated text categorization. In: Text mining and its applications. Springer, Berlin, pp 81–97

Dharwal P, Choudhury T, Mittal R, Kumar P (2017) Automatic sarcasm detection using feature selection. Paper presented at the 2017 3rd international conference on applied and theoretical computing and communication technology (iCATccT)
Dictionary C (2008) Cambridge advanced learner's dictionary: PONS-Worterbucher. Klett Ernst Verlag GmbH, Stuttgart
Dictionary ME, Rundell M (2007) Macmillan English dictionary. Macmillan Education, London
Dumais S, Chen H (2000) Hierarchical classification of web content. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, pp 256–263. ACM Press
Eke CI, Norman AA, Shuib L, Nweke HF (2019) A survey of user profiling: state-of-the-art, challenges, and solutions. IEEE Access 7:144907–144924. https://doi.org/10.1109/ACCESS.2019.2944243
Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15(1):3133–3181
Fersini E, Pozzi FA, Messina E (2015) Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. Paper presented at the 2015 IEEE international conference on data science and advanced analytics (DSAA)
Filatova E (2012) Irony and sarcasm: corpus generation and analysis using crowdsourcing. Paper presented at the LREC
Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3(Mar):1289–1305
Ghosh A, Veale T (2016) Fracking sarcasm using neural network. Paper presented at the proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis
Ghosh D, Guo W, Muresan S (2015) Sarcastic or not: word embeddings to predict the literal or sarcastic meaning of words. Paper presented at the proceedings of the 2015 conference on empirical methods in natural language processing
González-Ibánez R, Muresan S, Wacholder N (2011) Identifying sarcasm in Twitter: a closer look. Paper presented at the proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol 2
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182
Hall MA, Smith LA (1998) Practical feature subset selection for machine learning. pp 181–191
He X, Xu S (2010) Process neural networks: theory and applications. Springer, Berlin
Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification technical report department of computer science and information engineering. National Taiwan University, Taipei
Joshi A, Tripathi V, Patel K, Bhattacharyya P, Carman M (2016) Are word embedding-based features useful for sarcasm detection? arXiv preprint arXiv:1610.00883
Joshi A, Bhattacharyya P, Carman MJ (2017) Automatic sarcasm detection: a survey. ACM Comput Surv CSUR 50(5):73
Khattri A, Joshi A, Bhattacharyya P, Carman M (2015) Your sentiment precedes you: using an author's historical tweets to predict sarcasm. Paper presented at the proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis
Khodak M, Saunshi N, Vodrahalli K (2017) A large self-annotated corpus for sarcasm. arXiv preprint arXiv:1704.05579
Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering—a systematic literature review. Inf Softw Technol 51(1):7–15
Kumar HK, Harish B (2018) Sarcasm classification: a novel approach by using content based feature selection method. Proc Comput Sci 143:378–386
Kumar A, Sangwan SR, Arora A, Nayyar A, Abdel-Basset M (2019) Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access 7:23319–23328
Kunneman F, Liebrecht C, Van Mulken M, Van den Bosch A (2015) Signaling sarcasm: from hyperbole to hashtag. Inf Process Manage 51(4):500–509
Lee H-S, Lee H-R, Park J-U, Han Y-S (2018) An abusive text detection system based on enhanced abusive and non-abusive word lists. Decis Support Syst 113:22–31. https://doi.org/10.1016/j.dss.2018.06.009
Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22
Liebrecht C, Kunneman F, van Den Bosch A (2013) The perfect solution for detecting sarcasm in tweets# not. In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 29–37
Ling J, Klinger R (2016) An empirical, quantitative analysis of the differences between sarcasm and irony. Paper presented at the European semantic web conference

Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167
Liu P, Chen W, Ou G, Wang T, Yang D, Lei K (2014) Sarcasm detection in social media based on imbalanced classification. In: International conference on web-age information management. Springer, Cham, pp 459–471
Lunando E, Purwarianti A (2013) Indonesian social media sentiment analysis with sarcasm detection. In: 2013 international conference onadvanced computer science and information systems (ICACSIS). IEEE, pp 195–198
Manjusha P, Raseek C (2018) Convolutional neural network based simile classification system. Paper presented at the 2018 international conference on emerging trends and innovations in engineering and technological research (ICETIETR)
Manohar MY, Kulkarni P (2017) Improvement sarcasm analysis using NLP and corpus based approach. Paper presented at the 2017 international conference on intelligent computing and control systems (ICICCS)
McCallum A, Nigam K (1998) A comparison of event models for naive Bayes text classification. Paper presented at the AAAI-98 workshop on learning for text categorization
Mehndiratta P, Sachdeva S, Soni D (2017) Detection of sarcasm in text data using deep convolutional neural networks. Scalable Comput Pract Exp 18(3):219–228
Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. MIT Press, Cambridge
Mujtaba G, Shuib L, Raj RG, Majeed N, Al-Garadi MA (2017) Email classification research trends: review and open issues. IEEE Access 5:9044–9064
Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K et al (2018) Clinical text classification research trends: systematic literature review and open issues. Expert Syst Appl 116:494–520
Mukherjee S, Bala PK (2017a) Detecting sarcasm in customer tweets: an NLP based approach. Ind Manag Data Syst 117(6):1109–1126
Mukherjee S, Bala PK (2017b) Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. Technol Soc 48:19–27. https://doi.org/10.1016/j.techsoc.2016.10.003
Muresan S, Gonzalez-Ibanez R, Ghosh D, Wacholder N (2016) Identification of nonliteral language in social media: a case study on sarcasm. J Assoc Inf Sci Technol 67(11):2725–2737
Nithya K, Kalaivaani PD, Thangarajan R (2012) An enhanced data mining model for text classification. Paper presented at the 2012 international conference on computing, communication and applications (ICCCA)
Nweke HF, Teh YW, Al-Garadi MA, Alo UR (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl 105:233–261
Patro S, Sahu KK (2015) Normalization: a preprocessing stage. arXiv preprint arXiv:1503.06462
Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015
Provost FJ, Fawcett T (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. Paper presented at the KDD
Provost FJ, Fawcett T, Kohavi R (1998) The case against accuracy estimation for comparing induction algorithms. Paper presented at the ICML
Ptáček T, Habernal I, Hong J (2014) Sarcasm detection on Czech and English twitter. Paper presented at the proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers
Quinlan JR (1990) Decision trees and decision-making. IEEE Trans Syst Man Cybern 20(2):339–346
Rajadesingan A, Zafarani R, Liu H (2015a) Sarcasm detection on Twitter. Paper presented at the proceedings of the eighth ACM international conference on web search and data mining—WSDM '15
Rajadesingan A, Zafarani R, Liu H (2015b) Sarcasm detection on twitter: a behavioral modeling approach. Paper presented at the proceedings of the eighth ACM international conference on web search and data mining
Ramos J (2003) Using TF-IDF to determine word relevance in document queries. Paper presented at the proceedings of the first instructional conference on machine learning
Ranjan P, Yadav J, Saha S (2017) Proposed approach for sarcasm detection in Twitter. Indian J Sci Technol 10(25):1–8. https://doi.org/10.17485/ijst/2017/v10i25/114443
Rennie JD, Shih L, Teevan J, Karger DR (2003) Tackling the poor assumptions of naive Bayes text classifiers. Paper presented at the proceedings of the 20th international conference on machine learning (ICML-03)
Reyes A, Rosso P, Buscaldi D (2012) From humor recognition to irony detection: the figurative language of social media. Data Knowl Eng 74:1–12

Reyes A, Rosso P, Veale T (2013) A multidimensional approach for detecting irony in twitter. Lang Resour Eval 47(1):239–268
Riloff E, Qadir A, Surve P, De Silva L, Gilbert N, Huang R (2013) Sarcasm as contrast between a positive sentiment and negative situation. Paper presented at the proceedings of the 2013 conference on empirical methods in natural language processing
Saha S, Yadav J, Ranjan P (2017) Proposed approach for sarcasm detection in twitter. Indian J Sci Technol 10:25
Sahami M, Dumais S, Heckerman D, Horvitz E (1998) A Bayesian approach to filtering junk e-mail. Paper presented at the learning for text categorization: papers from the 1998 workshop
Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manage 24(5):513–523
Salton G, McGill MJ (1986) Introduction to modern information retrieval. Facet Publishing, London
Samonte MJC, Dollete CJT, Capanas PMM, Flores MLC, Soriano CB (2018) Sentence-level sarcasm detection in English and Filipino tweets. Paper presented at the Proceedings of the 4th international conference on industrial and business engineering—ICIBE' 18. http://delivery.acm.org/10.1145/3290000/3288172/p181-Samonte.pdf?ip=103.18.0.19&id=3288172&acc=ACTIVE%20SERVICE&key=69AF3716A20387ED%2EE7759EC8BE158239%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1562041412_216ad611ed7438dea30eb1738af6b7df. Accessed 24 Oct 2018
Schifanella R, de Juan P, Tetreault J, Cao L (2016) Detecting sarcasm in multimodal social platforms. Paper presented at the proceedings of the 2016 ACM on multimedia conference
Sintsova V, Pu P (2016) Dystemo. ACM Trans Intell Syst Technol 8(1):1–22. https://doi.org/10.1145/2912147
Sreelakshmi K, Rafeeque P (2018) An effective approach for detection of sarcasm in tweets. Paper presented at the 2018 international CET conference on control, communication, and computing (IC4)
Strapparava C, Valitutti A (2004) Wordnet affect: an affective extension of wordnet. Paper presented at the LREC
Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F (2017) Natural language processing based features for sarcasm detection: an investigation using bilingual social media texts. Paper presented at the 2017 8th international conference on information technology (ICIT)
Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F (2018) Mechanism for sarcasm detection and classification in malay social media. Adv Sci Lett 24(2):1388–1392
Suhaimin MSM, Hijazi MHA, Alfred R, Coenen F (2019) Modified framework for sarcasm detection and classification in sentiment analysis. Indones J Electr Eng Comput Sci 13(3):1175–1183
Sulis E, Farías DIH, Rosso P, Patti V, Ruffo G (2016) Figurative messages and affect in Twitter: differences between# irony,# sarcasm and# not. Knowl-Based Syst 108:132–143
Tsur O, Rappoport A (2012) What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. Paper presented at the proceedings of the fifth ACM international conference on web search and data mining
Tsur O, Davidov D, Rappoport A (2010) ICWSM—a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews. Paper presented at the fourth international AAAI conference on weblogs and social media
van der Aalst WM (2001) Exterminating the dynamic change bug: a concrete approach to support workflow change. Inf Syst Front 3(3):297–317
Wang Z, Wu Z, Wang R, Ren Y (2015) Twitter sarcasm detection exploiting a context-based model. Paper presented at the international conference on web information systems engineering
Wicana SG, İbisoglu TY, Yavanoglu U (2017) A review on sarcasm detection from machine-learning perspective. Paper presented at the 2017 IEEE 11th international conference on semantic computing (ICSC)
Yang M-S (1993) A survey of fuzzy clustering. Math Comput Model 18(11):1–16
Yang Y (1999) An evaluation of statistical approaches to text categorization. Inf Retrieval 1(1–2):69–90
Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. Paper presented at the ICML
Yao X (1999) Evolving artificial neural networks. Proc IEEE 87(9):1423–1447
Yavanoglu U, Ibisoglu TY, Wıcana SG (2018) Technical review: sarcasm detection algorithms. Int J Semant Comput 12(03):457–478
Yee Liau B, Pei Tan P (2014) Gaining customer knowledge in low cost airlines through text mining. Ind Manag Data Syst 114(9):1344–1359

Zhang M, Zhang Y, Fu G (2016) Tweet sarcasm detection using deep neural network. Paper presented at the proceedings of COLING 2016, The 26th international conference on computational linguistics: technical papers

Affiliations

Christopher Ifeanyi Eke[1,2] · Azah Anir Norman[1] · Liyana Shuib[1] · Henry Friday Nweke[1,3]

Liyana Shuib
liyanashuib@um.edu.my

Henry Friday Nweke
henry.nweke@ebsu.edu.ng

[1] Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Department of Computer Science, Faculty of Science, Federal University, P.M.B 146, Lafia, Nasarawa State, Nigeria

[3] Computer Science Department, Ebonyi State University, P.M.B 053, Abakaliki, Ebonyi State, Nigeria