Center for Open Access in Science ▪ https://www.centerprode.com/ojit.html
*Open Journal for Information Technology*, 2022, *5*(1), 1-16.
ISSN (Online) 2620-0627 ▪ https://doi.org/10.32591/coas.ojit.0501.010010

# Machine Learning Techniques, Features, Datasets, and Algorithm Performance Parameters for Sentiment Analysis: A Systematic Review

Bernard Ondara, Stephen Waithaka & John Kandiri

*Kenyatta University, Nairobi, KENYA*
*School of Engineering and Technology*

Lawrence Muchemi

*University of Nairobi, KENYA*
*Department of Computing and Informatics*

Received: 29 November 2021 ▪ Revised: 20 February 2022 ▪ Accepted: 28 March 2022

*Abstract*

The purpose of this paper is to review various studies on current machine learning techniques used in sentiment analysis with the primary focus on finding the most suitable combinations of the techniques, datasets, data features, and algorithm performance parameters used in most applications. To accomplish this, we performed a systematic review of 24 articles published between 2013 and 2020 covering machine learning techniques for sentiment analysis. The review shows that Support Vector Machine as well as Naïve Bayes techniques are the most popular machine learning techniques; word stem and n-grams are the most extensively applied features, and the Twitter dataset is the most predominant. This review further revealed that machine learning algorithms' performance depends on many factors, including the dataset, extracted features, and size of data used. Accuracy is the most commonly used algorithm performance metric. These findings offer important information for researchers and businesses to use when selecting suitable techniques, features, and datasets for sentiment analysis for various business applications such as brand reputation monitoring.

*Keywords*: sentiment analysis; machine learning technique; machine learning algorithm; sentiment classification technique; sentiment classification algorithm.

## 1. Introduction

The proliferation of Internet and Mobile technologies has led to the rapid adoption of social networks and micro-blogging sites. Equally, there is a rising trend in sharing user views and experiences about products and services daily on the Internet. For example, when customers want to purchase a new product or service, they often check on the reviews left behind by previous customers. By the same token, companies can access product and service reviews made by their clients through the different micro-blogging and social media networks like Twitter, Instagram, Facebook, YouTube, and LinkedIn to help them discover and eventually make improvements on the targeted products and/or services and/or make better business decisions (Ahmad et al., 2018). It is not feasible for humans to read all the posts (tweets, comments, and reviews) made by clients.


**Correspondence** Bernard Ondara, Kenyatta University, School of Engineering and Technology, Nairobi, KENYA. E-mail: ondara.Bernard@ku.ac.ke.

Consequently, this has become a fertile area for research, particularly in the design of automated computer algorithmic techniques for carrying out two popular and important tasks: text classification and sentiment analysis.

- Naive Bayes and Support Vector Machines are the two most popularly used machine learning techniques in sentiment analysis.
- Twitter data and reviews data are the most popular datasets used in sentiment analysis with machine learning techniques.
- The most popular features in sentiment analysis using machine learning techniques are Word stem, n-grams, and Bag of Word.
- Accuracy is the most commonly used algorithm performance metric in sentiment classification where machine learning techniques are used.

Sentiment analysis involves classifying a particular text into positive, neutral, or negative classes. From published literature, there exist three approaches to sentiment analysis. These include Lexicon-based, machine learning, and the blended technique called ensemble/hybrid approach, which bears some aspects from the other two approaches. Various Lexicon-based sentiment analysis techniques and tools have been explored (Ahmad et al., 2017b). Diverse machine learning tools and techniques for performing sentiment analysis have been explored and discussed in depth (Ahmad et al., 2017). In an attempt to improve sentiment classification and general sentiment analysis quality and accuracy, research has advanced to involve the combination of lexicon-based and machine-learning-based techniques in what is classically known as ensemble or hybrid techniques (Ahmad et al., 2017a). In the machine learning-based approach, several techniques have been used, including the following: Support Vector Machine (SVM), Naïve Bayes (NB), Random Forests (RF), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Decision Trees (DT), Logistic Regression, and Maximum Entropy (ME).

These algorithms generally belong to a class of algorithms called classification algorithms. These algorithms also fall in the group of supervised machine learning algorithms. The algorithms must first be trained using a pre-identified classes of outputs called training data and thereafter gain the capability of classifying real input data called test data. For purposes of sentiment analysis and text classification, many annotated datasets exist for different application domains. These datasets include Twitter data, Amazon product review dataset (Jindal et al., 2008), gender classification dataset (Mukherjee & Liu, 2010), and customer review dataset (Hu & Liu, 2004), among others. This study examined 235 papers on sentiment analysis utilizing numerous machine learning techniques published from 2013 to 2020. We especially used three online libraries: Science Direct, Academia.edu, and Research-gate. Based on the specific query strings used, we identified 235 articles. Upon applying the systematic review framework outlined in sections 3.3 and 3.4 of this study, 24 papers that met our inclusion criteria were selected for a thorough and comprehensive review.

This systematic review paper is structured in the following fashion. Section 2 defines the related works in this research sphere. Section 3 outlines the research methodology employed in this study. Section 4 gives a detailed review of the carefully selected papers. Section 5 is the discussion of the key results of this comprehensive review. Lastly, section 6 concludes this review paper.

## 2. Related work

Sentiment analysis is a research area that involves analyzing the sentiments, opinions, attitudes, emotions, appraisals, and evaluations of people towards entities like products, services, people, topics, brands, and their attributes (Liu, 2015). An opinion has the following four attributes:

- Object – the target of the opinion;
- Aspect – an object's targeted attribute;
- Sentiment orientation – what indicates if the opinion is positive, neutral, or negative;
- Opinion holder – this is the individual or party that articulates an opinion.

Considering the above attributes of an opinion (quintuple), sentiment analysis presents a very thought-provoking research area with multifaceted tasks (Kharde & Sonawane, 2016). Sentiment classification, subjectivity classification, aspect extraction, spam detection, and lexicon creation are some of the most frequently studied sentiment analysis tasks.

### 2.1 *Granularity levels in sentiment analysis*

Three sentiment analysis levels exist. They include aspect-level together with sentence-level and document-level (Saranya et al., 2016). At the aspect level, the primary objective is to discover the four attributes of the opinion. Aspect extraction and the subsequent aspect sentiment classification are the two primary tasks defined at this finer-grained level of sentiment analysis. A positive, neutral, or negative opinion on a given object is referenced to an object's attribute and not the entire object.

At the sentence level, the focus is to identify if a sentence carries a sentiment or not, in addition to assessing the sentimentality of particular independent sentences. This granularity is a bit more puzzling because the sentiment orientation is exceedingly reliant on the context in which the word is used. Common challenges in this level include handling sarcasm and comparisons at the sentence level.

The last level is the document-level granularity. In this case, the text is treated as one unit and consequently assigned a positive, neutral, or negative sentiment class. This involves assuming the whole document presents a single opinion about the opinion holder and cannot, therefore, be used where a document compares or evaluates multiple entities.

### 2.2 *Approaches to sentiment analysis*

Typically, sentiment analysis through machine learning techniques is performed using any of the following approaches or their combinations: unsupervised, supervised, and hybrid approaches. According to Boudad et al. (2018), the supervised machine learning approach uses algorithms like Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN) and Artificial Neural Networks (ANN). A huge dataset of categorized data is used to train a sentiment classifier or a set of multiple sentiment classifiers in this approach. The purpose of this is to develop a sentiment classification model that can predict the opinion or sentiment in new pieces of text. This is because oftentimes, labeled data is not available, or the process of annotating data is cumbersome.

On the other hand, unsupervised machine learning approaches use words together with their respective sentiments. Every word found in the lexicon bears sentiment polarity scores

that signify if the sentiment is positive, neutral, or negative. Lexicons can be created from a current corpus or dictionary. Contrasting this with the supervised approach, the unsupervised machine learning approach does not necessitate using labeled datasets. In its place, it requires an extensive lexicon that includes the largest possible number of sentiment words.

Instead of utilizing either the supervised or unsupervised approach, a few researchers have opted to integrate both of these approaches. The resulting new model, usually known as the semi-supervised approach, employs a huge quantity of unlabeled data and a small quantity of partially labeled data in creating improved sentiment classifiers. Then, the model classifies the unlabeled data employing supervised classifiers that have been trained using labeled data.

### 2.3 *Common sentiment classifiers / classification techniques*

In this section, we briefly describe some of the most popular machine learning techniques used in sentiment analysis. This list is therefore not exhaustive.

#### 2.3.1 *Naïve Bayes (NB)*

This sentiment classification technique is probabilistic and often performs well when large datasets are used. The classifier must first compute a posterior probability and then assume that the features involved are conditionally independent. Even so, to do away with unwanted effects, smoothing techniques are employed (Kaur, 2016).

#### 2.3.2 *Support Vector Machine (SVM)*

Like the NB model, this model is also probabilistic and must have training data for model training. SVM employs nonlinear mapping to locate a big margin between various classes. The classifier tries to discover a decision margin that makes the most of the separation gap between two classes. SVM classifier is highly accurate even though it requires additional time in model training. Contrasting with the NB classifier, SVM does not make any assumption regarding the conditional independence of classes (Ankit & Saleena, 2018).

#### 2.3.3 *Logistic Regression (LR)*

This model is employed in classification tasks. It is normally used to associate one definite independent variable with at least one independent variable. The LR classifier tries to identify a hyper-plane that makes the most of the separation gap amid classes (Ankit & Saleena, 2018).

#### 2.3.4 *Random Forests (RF)*

The RF classifier is an ensemble of decision-tree based classifiers. As such, the RF classifier creates a set of decision trees from the existing dataset for training. Upon getting votes from the various decision trees, the RF classifier determines the ultimate class or label of an object in the test dataset (Ankit & Saleena, 2018).

### 2.4 *Datasets*

The following are some of the common Twitter datasets.

2.4.1 *Twitter sentiment analysis dataset*

This particular dataset contains 99,989 tweets for model training. Each of the tweets is labeled as positive or negative. Of this set, 56,457 tweets are labeled as positive, while the remaining 43532 tweets are labeled as negative, according to Ankit and Saleena (2018).

2.4.2 *Stanford-sentiment140 corpus*

This dataset has 1.6 million tweets for model training. This set is evenly divided into positive and negative tweets (Go et al., 2009).

2.4.3 *First GOP debate Twitter sentiment dataset*

Crowdflower hosts this dataset, and it contains tweets about the first GOP debate held for the presidential nomination in 2016. The dataset has 13,871 tweets, with 2,236 being positive, 8,493 being negative, and the remaining 3,142 tweets bearing a neutral sentiment label (Ankit & Saleena, 2018).

2.4.4 *Health care reform (HCR)*

This dataset was collected from Twitter using the #hrc search tag (Go et al., 2009). It has 888 tweets, of which 365 are positive, and 523 are negative.

2.5 *Algorithm performance parameters*

Based on the review conducted, the performance of different sentiment classifiers is based on the following algorithm performance evaluation metrics: Accuracy, Recall, Precision, and F-score, whose formulae are given below (Ahuja et al., 2019). However, it was observed that most studies focus on accuracy alone while it is our understanding that the other three parameters could be used to give a better estimation of the performance of sentiment classification techniques.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Accuracy measures the percentage of correctly classified cases. Precision measures that ratio positive cases to the sum of all positive cases. Recall measures the number of correctly cases classified into a class in relation to all cases present. F-score is a measure of the weighted average of recall and precision. Where classes are unevenly distributed, this parameter is the best to use in gauging algorithm performance (Ahuja et al., 2019).

## 3. Research methodology

This study was carried out to obtain key information from the most current research publications on sentiment analysis with particular emphasis on machine learning techniques, features, datasets and algorithm performance metrics used. The articles are published in the last 6-7 years.

Systematic reviews scrutinize gaps amid various researches over some time (Kitchenham et al., 2009). A research methodology outlines the step-wise structure followed in carrying out the study. According to Brereton et al. (2007), an in-depth procedure, specific structure, and boundary lines should be used to guide the process of selecting the most pertinent study articles that bear the utmost quality. The methodology adopted in preparing this systematic review is provided in this section. The exact steps followed consist of searching for relevant publications, publication inclusion criteria, publication exclusion criteria, and the findings in the publications. A Systematic Literature Review Guidelines document in the domain of software engineering was equally important in this regard (Ashraf & Aftab, 2017; Ashraf, 2017; Anwer & Aftab, 2017).

### 3.1 *Research questions*

To reflect the primary objectives of this study, especially while carrying out the systematic review of the selected articles, five research questions required to be answered:

RQ1: Which current machine learning techniques are the most popular in sentiment analysis?

RQ2: What machine learning techniques are particularly applied in Twitter sentiment classification?

RQ3: What kinds of datasets are commonly used for performance evaluation?

RQ4: What are the main features of the datasets for Twitter sentiment analysis?

RQ5: What are the key algorithm performance metrics for sentiment classifiers?

### 3.2 *Query string and search space*

A query string refers to the blend of certain keywords applied to extract search publications or articles from the chosen online libraries. The specific keywords that were extracted from the six research questions include: Machine learning techniques, Sentiment analysis, Twitter sentiment analysis, machine learning techniques. The following search query was completed using the above keywords:

> ((“Machine learning techniques” OR “Machine learning algorithms”) AND (“Twitter sentiment analysis”))

The following popular online search libraries were mainly used: Science Direct, Research Gate, and Academia.edu. Since these libraries differ in search characteristics, we adjusted the query strings slightly to ensure that more appropriate articles are extracted. This means that the same query was applied at different times, with each iteration having some slight modifications to the keywords’ arrangement.

3.3 *Selection criteria*

This section involves articles Inclusion Criteria (IC) rules:

IC1: Articles published from 2013 to 2020;

IC2: Articles that used one machine learning technique in sentiment analysis;

IC3: Articles that used hybrid techniques that involved various machine learning algorithms;

IC4: Articles with experiments and results;

IC5: Articles written in English;

Besides the inclusion above, only those articles bore greater relevance to the research questions considered for this study.

3.4 *Quality assessment*

For purposes of achieving high-quality results, the following specific quality assessment parameters were considered:

- Credible electronic scientific libraries were used to extract the appropriate research materials
- To ascertain the highest quality of results, only the most recent publications were considered
- The process of selecting the articles was unbiased.
- The systematic literature review steps discussed above were strictly followed.

3.5 *Data extraction and synthesis*

Upon completing the search process, 24 most relevant publications were identified.

4. Literature analysis

The selected papers were analyzed and the results presented in the tables below. The results are presented based on the machine learning technique(s) the studies used, features extracted, highest accuracy attained, and the article's reference. The reason for this breakdown is to help identify what machine learning techniques, features, and algorithm performance parameters are used for different datasets and social media platforms. It makes it easier to analyze, for instance, the techniques, features and accuracy levels of the techniques used for different social media platforms like Twitter and Facebook.

4.1 *Synthesis of the experimental results*

Tables 1-6 represent sentiment classification techniques using machine learning approach.

Table 1. Twitter dataset

| Machine Learning Technique | Features Extracted | Accuracy | Reference |
|---|---|---|---|
| SVM, Back-Propagation Neural Networks (BPNN), NB, Decision Tree | Word stem | 96.06% | Hammad & Mouhammd, (2016) |
| SVM, NB | Word stem | 95% F-score | Duwairi (2015) |
| NB, DT | Word stem | 64.85% | Al-Horaibi & Khan (2016) |
| SVM, NB, KNN | Word stem, n-grams | 67% | Duwairi & El-Orfali, 2014) |
| SVM, NB, KNN | Word stem, n-grams | 69.97% | Duwairi & Qarqaz (2014) |
| NB, SVM, MaxEnt, ANN | Unigram, Bigram, Hybrid of Unigram and Bigram | 92% for SVM with PCA | Anjaria & Guddeti (2014) |
| MNB, SVM, BNB, Passive Aggression KNN, LR, SGD | n-grams | 69.1% | Nabil et al. (2015) |
| Prind, KNN, NB, SVM, RF, NBMN (NB Multinomial) | Text Blob, SentiWordNet, and WSD | 79% for NB with WSD; 76% for NB with TextBlob | Hasan et al. (2018) |
| SVM | Sentence-level features, Standard features, linguistic | 95% | Ibrahim et al. (2015) |
| Complement NB | Sentiment Lexicon features, text-related features, emoticon-based features | 79.4% | El-Beltagy & Ali (2013) |
| NN, Hierarchical and DBSCAN clustering | | | Stojanovski et al. (2016) |
| RF | User and network features | *Improved* | Novalita et al. (2019) |

Table 2. Facebook dataset

| Machine Learning Technique | Features Extracted | Accuracy | Reference |
|---|---|---|---|
| SVM, KNN, BN | Word stem + n-grams | 69.97% | Duwairi & Qarqaz (2016) |
| SVM, Back-Propagation Neural Networks (BPNN), NB, Decision Tree | Word stem | 96.06% | Hammad & Mouhammd (2016) |

Table 3. YouTube dataset

| Machine Learning Technique | Features Extracted | Accuracy | Reference |
|---|---|---|---|
| NB, SVM, DT, based lexicon | Unigram, bigram, word stem, Sentiment orientation weight | 95% | Elawady et al. (2015) |
| SVM, Back-Propagation Neural Networks (BPNN), NB, Decision Tree | Word stem | 96.06% | Hammad & Mouhammd (2016) |

Table 4. Reviews datasets: Amazon reviews, hotel reviews, book reviews, multi-domain reviews

| Machine Learning Technique | Features Extracted | Accuracy | Reference |
|---|---|---|---|
| NB | Word features | 75% | Jain et al. (2016) |
| SVM, ANN, and MaxEnt | Opinion features, stylistic features, discourse markers, morpholexical features, and domain-dependent features | 85.06% | Bayoudhi et al. (2015) |
| Linear SVM, LR, Bernoulli NB, KNN, Stochastic Gradient Descent. | n-grams, Lexicon entries | 88% | ElSahar & El-Beltagy (2015) |
| SVM | Low-level stem | 83% | Cherif et al. (2015) |
| Hierarchical Classification using: SVM, KNN, DT, NB | Bag-Of-Word | 57.8% | Al Shboul et al. (2015) |
| SVM | Standard features, sentence-level features, linguistic features, | 95% | Ibrahim et al. (2015) |
| SVM, NB, Back-Propagation Neural Networks (BPNN), Decision Tree | Word stem | 96.06% | (Hammad & Mouhammd (2016) |

Table 5. Aljazeera dataset

| Machine Learning Technique | Features Extracted | Accuracy | Reference |
|---|---|---|---|
| SVM, KNN, NB | Word stem + n-grams | 96% | Duwairi & El-Orfali (2014) |
| SVM, MaxEnt, and ANN | Opinion features, domain-dependent features, discourse markers, morpholexical Features and stylistic features | 85.06% | Bayoudhi et al. (2015) |

Table 6. Other datasets: Blogs, Goodreads.com, Subjectivity, NetvizzApp

| Machine Learning Technique | Features Extracted | Accuracy | Reference |
|---|---|---|---|
| SVM, NB | Word stem + n-grams | 75% | Akaichi et al. (2013) |
| J48, Decision Table, KNN, SVM, NB, MNB | Word stem | 58% | Al Shboul et al. (2015) |
| NB, Variational Expectation-Maximization (VEM) | | 84.6% | Adeborna & Siau (2014) |
| KNN | Link, photo, status update, and video. | 82.3% | Poecze et al. (2018) |

### 4.2 *Accuracy of the techniques/algorithms*

In their study, Ibrahim et al. (2015) used SVM alone on Twitter reviews and comments, attaining a 95% accuracy with an extensive feature set. A different study by Cherif et al. (2015) applied SVM to hotel reviews and attained 83% accuracy with feature weights generated using a new mathematical approach. This significant variation in the accuracy shows that SVM accuracy

depends on the dataset and features used. In other cases, SVM was used in combination with other machine learning methods. For instance, Duwairi and Al-Rifai (2015) used an SVM and NB ensemble on the Twitter dataset and attained 87% F-measure accuracy.

The lowest SVM performance was recorded by Al Shboul et al. (2015), who applied it to the goodreads.com dataset and attained a 58% accuracy but with using multi-way classification where other algorithms such as KNN, MB, MNB, Decision Table, and J48 were also studied. This is a classic example of where ensemble methods perform lower than independent methods. The highest accuracy was obtained where SVM was used in the study by Hammad and Mouhammd (2016), who used SVM, BPNN, NB, and Decision Trees on various datasets, and this resulted in 96.06% accuracy with POS tagging.

In the Naive Bayes (NB) method, the highest accuracy recorded in the papers reviewed is also in the same study where SVM performed the best according to Hammad and Mouhammd (2016). NB was used alone, such as in the study by Duwairi and Qarqaz (2014) using Amazon reviews dataset and word features, the accuracy was 75%. A variation of NB, Complement Naïve Bayes, was applied in a study by ElSahar and El-Beltagy (2015) with an extensive set of Twitter features that resulted in 79.4% accuracy. The lowest recorded performance of NB in this study by Bayoudhi et al. (2015). Just as was the case with SVM, this is one reason that many studies argue that SVM and NB have comparatively similar accuracy levels.

It is worth noting that other machine learning methods have good accuracy levels even though they are not popular. For instance, the Random Forest (RF) method achieved 84.0% accuracy, while the KNN method achieved 82.3% accuracy in a study by Poecze et al. (2018). Unfortunately, other techniques such as Decision Trees, Logistic Regression, and Maximum Entropy were not used alone in any of the 24 studies; hence difficult to get their independent accuracies.

### 4.3 *Datasets used*

13 out of the 24 papers reviewed used Twitter as a source of data. Other studies used different data sources, including Aljazeera (Duwairi & El-Orfali, 2014), Facebook and Blogs (Akaichi et al., 2013), Amazon Reviews (Jain et al., 2016), goodreads.com (Al Shboul et al., 2015), Aljazeera movie reviews (Bayoudhi et al., 2015), YouTube comments (Elawady et al., 2015; Baccouche et al., 2019), Multi-domain reviews (ElSahar & El-Beltagy, 2015), hotel reviews (Cherif et al., 2015), book reviews (Al Shboul et al., 2015), subjectivity dataset (Hammad & Mouhammd, 2016; Poecze et al., 2018), and a combination of Hotels reviews, Facebook, Twitter, and YouTube (Hammad & Mouhammd, 2016).

### 4.4 *Features extracted*

Different studies used different features from their datasets. The popular feature, Word stem, was used in the studies by Duwairi and Al-Rifai (2015), Duwairi and El-Orfali (2014), Duwairi and Qarqaz (2016), Akaichi et al. (2013), Duwairi et al. (2015) and Al Shboul et al. (2015). Another common feature, n-gram, was used in the following studies: Duwairi and El-Orfali (2014), Duwairi and Qarqaz (2016), Akaichi et al. (2013), Nabil et al. (2015), ElSahar and El-Beltagy (2015), Duwairi et al. (2015), and Aldayel and Azmi (2015).

Some of the reviewed papers also used features that were not common or popular among the reviewed papers. These include Bag of Words (BoW), Bigrams, unigrams, word features, opinion features, Sentiment Lexicon features, sentence-level features, emoticon-based features, discourse markers, user and network features, stylistic features, text-related features, domain-dependent features, and morpholexical features, Text Blob, SentiWordNet, and WSD,

Sentiment orientation weight, Lexicon entries, Low-level stem, Standard features, linguistic features, and others.

## 5. Results and discussions

The discussion of our findings below is aligned with the research questions of this study.

*RQ1: Which current machine techniques are the most popular in sentiment analysis?*

The papers reviewed machine learning techniques used in sentiment analysis. Out of the 24 papers reviewed, 20 articles used Naïve Bayes (NB) technique in its simple form or other enhanced forms like Complement Naïve Bayes (CNB), Bernoulli NB (BNB), or Multinomial Naïve Bayes (MNB). This qualifies the Naïve Bayes technique as the most popular (83.3%) machine technique for sentiment analysis based on the papers reviewed. This was closely followed by Support Vector Machine (SVM) that was used in 16 of the 24 papers. This is the popularity of 66.67%. Interestingly, SVM was not used in other forms apart from a particular use of Linear SVM in one study.

The other machine methods attracted a usage popularity of less than 50%. For instance, K-Nearest Neighbor (KNN) method was used in 8 studies, which presents a 33.33% popularity among the considered studies. This was followed by Decision Trees that was used in 5 studies (20.83% popularity), Artificial Neural Networks (NN) was used in 4 studies (16.67% popularity), and Random Forests (RF) was used in 3 studies representing a 12.5% popularity. Both Maximum Entropy (ME) and Logistic Regression (LR) methods were used in 2 studies, representing a popularity of 8.33%.

*RQ2: What machine learning techniques are popularly applied to sentiment classification on Twitter?*

Of the 24 articles reviewed, 13 of them used Twitter datasets, which confirms the popularity of Twitter as a data source for sentiment analysis among popular studies. Interestingly, this review observed a positive correlation between machine learning techniques generally used for sentiment analysis and the those specifically used for Twitter sentiment analysis. In particular, this review observed that nine of the reviewed papers used SVM. Nine of the reviewed papers equally used the NB technique. Both algorithms represent a 69.23% popularity among researchers in Twitter sentiment analysis. This confirms the popularity of these two machine techniques among researchers in sentiment analysis. The other techniques in their decreasing popularity are KNN, ANN, RF, DT, MaxEnt, LR. The RF and DT techniques tied at 2 articles each while MaxEnt and LR tied at 1 article. This still confirms that SVM and NB are the most popularly or commonly used techniques for Twitter sentiment analysis.

*RQ3: What kinds of datasets are commonly used for performance evaluation?*

The 24 articles reviewed in this paper featured the application of different datasets from various platforms. This included Twitter, Facebook, YouTube, Amazon Reviews, Hotel Reviews, and Movie Reviews. Twitter social media network is the most popular as it offered the most popular dataset used in 12 of the studies, representing 50% popularity. This was followed by reviews (books, movies, hotels), representing 25% popularity of this study. Facebook was used in 3 studies, which is 12.5% in popularity, while YouTube had 8.33% popularity. This shows that Twitter is the most popular platform for studies on sentiment analysis. This is because Twitter data is easily accessible, available, and the quantity of tweets is good for model development, training, evaluation, and implementation.

*RQ4: What are the main features of the datasets for Twitter sentiment analysis?*

The papers reviewed contained machine learning techniques that exploited various features. The list of features used in the studies includes word stem, n-grams, unigrams and bigrams, hybrid unigrams and bigrams, word features, opinion features, Bag of Words (BoW), discourse markers, stylistic markers, morpho-lexical features, sentiment orientation weights, lexicon entries, low-level stem, domain-dependent features, and standard features among others. The most popular features in their order of decreasing popularity are Word stem, n-grams, and Bag of Word tying with unigrams/bigrams or a combination of unigrams with bigrams. The other features appeared only in one study each. The popularity of word stem was 33.33%, featuring in 8 of the 24 reviewed papers. In comparison, n-gram followed closely with a popularity of 20.83% as both n-grams, and unigram/bigram each scored 8.3% popularity.

*RQ5: What are the key algorithm performance metrics for sentiment classifiers?*

With reference to section 2.5, this study identified accuracy as the main and most popularly used measure of the performance of machine learning techniques in sentiment analysis. However, it should be noted that other parameters like Recall, Precision and F-score could be used to select better machine learning algorithms for different application areas in sentiment analysis. For example, if data classes are unevenly distributed, then it is best to use F-score parameter.

*Limitations of this study*

The following are the limitations of this study:

1. The performance of the machine learning algorithms reviewed was reported by various researchers who tested them. These researchers' actual performance or accuracy levels may be inaccurate, and this may generally affect the analysis of this research findings.

2. While a stringent approach was used in selecting the 24 papers reviewed in this study, there may be a few research works relevant to this study that was still left out. This may arise from a few factors, including but not limited to the inclusion criteria that were used in this study.

3. The reviewed articles showed varying performances of the same machine learning technique, making it challenging to pinpoint the best performing technique.

4. This review was limited to conventional machine learning techniques. There was no coverage of deep learning techniques.

This paper came up with five key findings that could contribute greatly to the advancement of sentiment analysis in the future. First, whereas different methods have been used to carry out a sentiment analysis on data from different social media platforms, this review establishes that the commonest machine learning techniques in the decreasing order of their popularity are: Naïve Bayes (NB), Support Vector Machines (SVM), and K-nearest neighbor (KNN), Decision Trees, Artificial Neural Networks (ANN) and Random Forests. Of these, the most popular three are NB, SVM, and KNN. Other techniques such as were used in one article and therefore considered very rare. It was noted that the choice of the technique was dependent on the data. NB and SVM have similar accuracy levels. Researchers' concern is to factor in text structure, data volume, and duration taken for model training and running. Machine learning techniques work well even with huge quantities of data, which requires more time to train the model than lexicon-based approaches. However, to improve the accuracy and quality of sentiment classification and analysis results, we suggest the application of ensemble approaches that involve combining machine learning methods with lexicon-based methods.

Second, this review shows that the commonest machine learning techniques used in carrying out sentiment analysis on different datasets are equally the commonest techniques used on Twitter datasets. They include NB, SVM, KNN, and ANN. This finding provides a ready recipe for those interested in studying these popular algorithms or even combining them to achieve better

results by creating hybrid models that benefit from the advantages of the techniques combined while diminishing their disadvantages or limitations when used independently. Moreover, these findings give researchers room to investigate unpopular machine learning techniques for sentiment analysis.

Third, we also identified Twitter as the commonest social media platform for extracting datasets suitable for sentiment analysis throughout this review. This is evident because most of the papers we reviewed used Twitter as their preferred data source. This is because Twitter is available, has rich content in the form of tweets, and is accessible. Checking on any common topic on Twitter could reveal that daily, there are millions of tweets. Conversely, we observed that there is little focus on other social networks like YouTube, Facebook, and WordPress blogs. Although the content and structure of data on these other social media platforms may differ from that on Twitter, this area is worth researching. It may yield new interesting findings and knowledge.

Fourth, this review shows that Word stem and n-grams are the commonest features used in sentiment analysis both on Twitter and other platforms or datasets. It would be interesting to investigate further if other features would yield better results than these common ones hence becoming a research area of interest.

Finally, accuracy is the single-most commonly referenced and used sentiment classifier performance parameter. In all the 24 papers reviewed, accuracy was considered and reported. One paper however simply indicated that the accuracy improved without giving the percentage.

## 6. Conclusion and future work

Sentiment analysis is currently a hot area of research within the larger knowledge discovery domain. Considering the huge volumes of data generated daily on the different social networking and micro-blogging sites in the form of tweets, posts, comments, and reviews, sentiment analysis techniques are often applied to get useful insights to help with brand reputation monitoring, getting the sentiments of the public regarding a given product or service just before it is launched, predicting election results, and a plethora of other applications. Three approaches to sentiment analysis are available. This systematic review presents research on sentiment analysis on data from social media networks and microblogging websites.

In conclusion, sentiment analysis has seen wide-ranging applications in various areas, including brand reputation monitoring, forecasting political election results, disaster location and response, data security awareness creation, business strategy and quality improvement, disease outbreak monitoring, and perceptions of people towards certain sports. These vast application areas show that sentiment analysis is useful in improving decision-making through gathering and analyzing people's perceptions towards a phenomenon, concept, person, or thing. We recommend that in the future, further studies be carried out to create a universal sentiment analysis model that could be applied to various data types and other social networking sites for purposes of obtaining user sentiments in a bid to expand the application of sentiment analysis in real life.

## Acknowledgements

This research did not receive any specific grant from funding agencies in the public commercial, or not-for-profit sectors.

The authors declare no competing interests.

## References


Adeborna, E., & Siau, K. (2014). An approach to sentiment analysis – The case of airline quality rating. *Proceedings - Pacific Asia Conference on Information Systems, PACIS 2014*, *363*.

Ahmad, M., Aftab, S., Ali, I., & Hameed, N. (2017a). Hybrid tools and techniques for sentiment analysis: A review. *International Journal of Multidisciplinary Sciences and Engineering*, *8*(4), 28-33.

Ahmad, M., Aftab, S., Ali, I., & Hameed, N. (2017b). Tools and techniques for lexicon driven sentiment analysis. *International Journal of Multidisciplinary Sciences and Engineering*, *8*(1), 17-23.

Ahmad, M., Aftab, S., Bashir, M. S., & Hameed, N. (2018). Sentiment analysis using SVM: A systematic literature review. *International Journal of Advanced Computer Science and Applications*, *9*(2), 182-188. https://doi.org/10.14569/IJACSA.2018.090226

Ahmad, M., Aftab, S., Muhammad, S., & Ahmad, S. (2017). Machine Learning Techniques for Sentiment Analysis: A Review. *Int. J. Multidiscip. Sci. Eng*, *8*(3), 27–32.

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, 341-348. https://doi.org/10.1016/j.procs.2019.05.008

Akaichi, J., Dhouioui, Z., & López, M. (2013). Social networks' text mining for sentiment classification: the case of Facebook'statuses updates. *17th International Conference*, 640-645.

Al-Horaibi, L., & Khan, M. B. (2016). Sentiment analysis of Arabic tweets using text mining techniques. *First International Workshop on Pattern Recognition*, *10011*(July 2016), 100111F. https://doi.org/10.1117/12.2242187

Al Shboul, B., Al-Ayyouby, M., & Jararwehy, Y. (2015). Multi-way sentiment classification of Arabic reviews. In *2015 6th International Conference on Information and Communication Systems, ICICS 2015* (pp. 206-211). https://doi.org/10.1109/IACS.2015.7103228

Aldayel, H. K., & Azmi, A. M. (2015). Arabic tweets sentiment analysis – A hybrid scheme. *Journal of Information Science*, *42*(6), 782-797. https://doi.org/10.1177/0165551515610513

Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. *2014 6th International Conference on Communication Systems and Networks, COMSNETS 2014*. https://doi.org/10.1109/COMSNETS.2014.6734907

Ankit, & Saleena, N. (2018). An ensemble classification system for Twitter sentiment analysis. *Procedia Computer Science*, *132*(Iccids), 937-946. https://doi.org/10.1016/j.procs.2018.05.109

Anwer, F., & Aftab, S. (2017). Latest customizations of XP: A systematic literature review. *International Journal of Modern Education and Computer Science*, *9*(12), 26-37. https://doi.org/10.5815/ijmecs.2017.12.04

Ashraf, S. (2017). Scrum with the spices of agile family: A systematic mapping. *International Journal of Modern Education and Computer Science*, *9*(11), 58-72. https://doi.org/10.5815/ijmecs.2017.11.07

Ashraf, S., & Aftab, S. (2017). Latest transformations in scrum: A state of the art review. *International Journal of Modern Education and Computer Science*, *9*(7), 12-22. https://doi.org/10.5815/ijmecs.2017.07.02

Baccouche, A., Garcia-Zapirain, B., & Elmaghraby, A. (2019). Annotation technique for health-related tweets sentiment analysis. *2018 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2018*, 382-387. https://doi.org/10.1109/ISSPIT.2018.8642685

Bayoudhi, A., Belguith, L. H., & Ghorbel, H. (2015). Sentiment classification of Arabic documents:

Experiments with multi-type features and ensemble algorithms. *29th Pacific Asia Conference on Language, Information and Computation, PACLIC 2015*, 196-205.

Boudad, N., Faizi, R., Oulad Haj Thami, R., & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal*, *9*(4), 2479-2490. https://doi.org/10.1016/j.asej.2017.04.007

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571-583. https://doi.org/10.1016/j.jss.2006.07.009

Cherif, W., Madani, A., & Kissi, M. (2015). A new modeling approach for Arabic opinion mining recognition. In *2015 Intelligent Systems and Computer Vision, ISCV 2015*. https://doi.org/10.1109/ISACV.2015.7105541

Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science*, *40*(4), 501-513. https://doi.org/10.1177/0165551514534143

Duwairi, R. M., Ahmed, N. A., & Al-Rifai, S. Y. (2015). Detecting sentiment embedded in Arabic social media - A lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*, *29*(1), 107-117. https://doi.org/10.3233/IFS-151574

Duwairi, Rehab M. (2015). Sentiment analysis for dialectical Arabic. *2015 6th International Conference on Information and Communication Systems, ICICS 2015*, 166-170. https://doi.org/10.1109/IACS.2015.7103221

Duwairi, Rehab M., & Qarqaz, I. (2014). Arabic sentiment analysis using supervised classification. *Proceedings – 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014*, *August*, 579-583. https://doi.org/10.1109/FiCloud.2014.100

Duwairi, Rehab M., & Qarqaz, I. (2016). A framework for Arabic sentiment analysis using supervised classification. *International Journal of Data Mining, Modelling and Management*, *8*(4), 369. https://doi.org/10.1504/ijdmmm.2016.10002311

El-Beltagy, S. R., & Ali, A. (2013). Open issues in the sentiment analysis of Arabic social media: A case study. *2013 9th International Conference on Innovations in Information Technology, IIT 2013*, 215-220. https://doi.org/10.1109/Innovations.2013.6544421

Elawady, R., Barakat, S., & Elrashidy, N. (2015). Sentimentanalysis for Arabic and English datasets. *International Journal of Intelligent Computing and Information Sciences*, *15*(1), 55-70. https://doi.org/10.21608/ijicis.2015.10911

ElSahar, H., & El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9042*, 23-34. https://doi.org/10.1007/978-3-319-18117-2_2

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 1-6.

Hammad, M., & Mouhammd, A. (2016). *Sentiment analysis for arabic reviews in social networks using machine learning*. *Apri*, 131-139. https://doi.org/10.1007/978-3-319-32467-8_13

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for Twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *KDD-2004 – Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *November*, 168-177. https://doi.org/10.1145/1014052.1014073

Ibrahim, H. S., Abdou, S. M., & Gheith, M. (2015). Sentiment analysis for modern standard Arabic and colloquial, ArXiv Prepr. *International Journal on Natural Language Computing (IJNLC)*,

*4*(2), 95-105. https://doi.org/10.1109/ReTIS.2015.7232904

Jain, J., Panchal, P., Suryawanshi, N., & Shinde, A. A. (2016). Sentiment analysis using supervised machine learning. *Imperial Journal of Interdisciplinary Research*, *2*(6), 2454-1362.

Jindal, N., Liu, B., & Street, S. M. (2008). Opinion spam and analysis. *Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008)*. https://doi.org/10.1145/1341531.1341560

Kaur, J. (2016). A review paper on Twitter sentiment analysis techniques. *International Journal for Research in Applied Science & Engineering Technology*, *4*(X), 61-70.

Kharde, V., & Sonawane, S. S. (2016). Sentiment analysis of Twitter data: A survey of techniques. *International Journal of Computer Applications*, *139*(11), 5-15. https://doi.org/10.5120/ijca2016908625

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, *51*(1), 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

Liu. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. In *Cambridge University Press*. https://doi.org/10.14569/ijacsa.2018.090981

Mukherjee, A., & Liu, B. (2010). Improving gender classification of blog authors. *EMNLP 2010 – Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 207-217.

Nabil, M., Aly, M., & Atiya, A. F. (2015). ASTD: Arabic sentiment tweets dataset. *Conference Proceedings – EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, *September*, 2515-2519. https://doi.org/10.18653/v1/d15-1299

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019). Cyberbullying identification on Twitter using random forest classifier. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012029

Poecze, F., Ebster, C., & Strauss, C. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Procedia Computer Science*, *130*, 660-666. https://doi.org/10.1016/j.procs.2018.04.117

Saranya, N., Phil, M., & Gunavathi, R. (2016). A study on various classification techniques for sentiment analysis on social networks. *International Research Journal of Engineering and Technology*, *3*(8), 1332-1337.

Stojanovski, D., Strezoski, G., Madjarov, G., & Dimitrovski, I. (2016). Finki at SemEval-2016 task 4: Deep learning architecture for Twitter sentiment analysis. *SemEval 2016 – 10th International Workshop on Semantic Evaluation, Proceedings*, 149-154. https://doi.org/10.18653/v1/s16-1022