Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

A Comprehensive Survey on Sentiment Analysis and Opinion Mining on Social Media Using Machine Learning Techniques

Prof. Divyashree G[1], Dr. Kamalakshi Naganna[2]
[1]Research Scholar, Department of Computer Science and Engineering,
Sapthagiri College of Engineering, Visvesvaraya Technological University, Belagavi-590018, India
[2]Head of the Department, Department of Computer Science and Engineering,
Sapthagiri College of Engineering, Visvesvaraya Technological University, Belagavi-590018, India.
[1]divyashreeraju@gmail.com, [2] kamalnags@gmail.com

Cite this paper as: Divyashree G, Kamalakshi Naganna (2024) A Comprehensive Survey on Sentiment Analysis and Opinion Mining on Social Media Using Machine Learning Techniques. *Frontiers in Health Informatics, 13* (3), 888-906.

Abstract

The growing usage of Internet-based applications, particularly social media platforms and blogs, resulted in an unprecedented flow of thoughts, reviews, and opinions. Sentiment Analysis (SA), also known as opinion mining, has emerged as a vital technique for systematically gathering and assessing people's sentiments, opinions, and impressions on a wide range of subjects, whether they are products, themes, or services, in this age of digital connectivity. The variety of public mood data has shown to be a valuable resource for corporations, governments, and individuals, enabling sound decision-making. However, impediments to precise judgment of sentiment polarity and precise interpretation of feelings develop as a result of SA implementation. SA is a sophisticated science that identifies and extracts subjective information from textual material. This is performed by utilizing strong Natural Language Processing (NLP) and text mining techniques. Our post aims to provide a thorough overview of the complicated procedures that underpin SA, as well as an examination of its problems. The SA process begins with data gathering from publically available datasets, followed by a number of data pre-processing activities such as converting text to lowercase, dealing with contractions, tokenizing, removing short and repeated words, and so on. Following that, feature extraction, including content-based, document-based, and texture-based features, is examined. Following that, the technique investigates Feature Selection (FS) approaches such as filter, wrapper, embedding, and hybrid procedures. Finally, we investigate numerous classification methods for sentiment detection, including machine learning, lexicons, and hybrid approaches. By presenting an in-depth review of key SA techniques and procedures, this study presents a comprehensive evaluation report that can serve as a solid foundation for future studies.

Keywords— *Sentiment Analysis, Natural Language Processing, Machine Learning, Review data, Accuracy*

Introduction

SA is the computer analysis of people's thoughts, feelings, and attitudes about a variety of themes, including services, goods, events, issues, and topics, as well as the traits connected with them [1, 2]. This survey provides useful information by measuring public opinion on specific problems. It also helps with the comprehension, interpretation, and prediction of social processes. SA is critical in business for making strategic decisions and knowing customer perceptions on products and services [3]. Understanding the consumer is crucial in today's customer-centric corporate landscape. The proliferation of online discussion forums, product review

websites, e-commerce, and social media has resulted in a constant flow of ideas and thoughts [4]. The sheer volume of this data poses a challenge for companies seeking to comprehend the collective sentiments and attitudes of their customers toward their offerings. However, the growth of internet-generated material, along with tools such as SA, allows marketers to obtain insight into customer sentiments. Extraction of feelings from product reviews, for example, enables marketers to identify clients who might need extra care, resulting in higher customer happiness, increased sales, and overall business profits. SA is an interdisciplinary field that draws on disciplines like sociology [5], psychology [6], NLP, and ML [7]. Recent advancements, driven by the exponential growth of data and computing power, have enabled more sophisticated forms of analysis. SA, a multidimensional field, encompasses various distinct tasks, each finely tuned to serve specific objectives in extracting insights related to sentiment from textual data [8]. Let's explore these SA tasks in greater depth with the help of Figure 1.

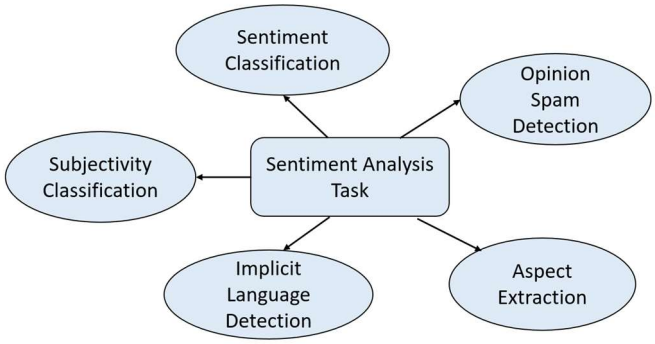

Fig. 1. Sentiment analysis task

- Subjectivity Classification: The classification of subjectivity focuses on determining whether a specific text transmits subjective ideas, feelings, or personal viewpoints or if it predominantly conveys objective, factual facts. This task is critical in settings such as news categorization, where distinguishing between subjective commentary and objective news reports is critical for effectively categorizing information.
- Sentiment Classification: Sentiment classification, also known as sentiment polarity classification, is the process of classifying text into predetermined sentiment classes such as positive, negative, or neutral. It provides useful insights into public opinion and is commonly used in sectors such as product evaluations and social media to understand how people perceive items, services, or events.
- Opinion Spam Detection: Opinion spam detection is critical for spotting false or misleading opinions, especially on online review platforms and forums. Its purpose is to screen out inauthentic reviews or comments intended to mislead users, hence preserving the credibility of online information sources [9].
- Aspect Extraction: Aspect extraction is concerned with recognizing certain characteristics or qualities stated in written information, most notably in product or service reviews. It provides a comprehensive view of what customers are talking about, allowing firms to focus on areas that need improvement and improve overall product or service quality [10].
- Implicit Language Detection: The goal of implicit language identification is to discover thoughts and ideas that are not directly expressed but are inferred from context, sarcasm, or suggested meanings. This duty is especially important in social media, where nuance and implicit language are prevalent. It helps sentiment analysers understand sentiment nuances hidden beneath the surface of textual data.

The review paper aims to cover every aspect of the sentiment analysis process, including data gathering,

processing, FS, classification, and the challenges and applications associated with it. The following is how the paper is organized: Section I provides an overview of sentiment analysis and its applications. Section II describes the ML-based sentiment analysis research process. Sections III and IV concentrate on data-collecting methodologies and data-cleaning procedures. In Sections V and VI, we look at different strategies for extracting essential features and selecting features that minimize data dimensions. Section VII delves into classifiers and an in-depth examination of prior work on sentiment analysis. Section VIII discusses the challenges encountered in the topic of sentiment analysis. Finally, the survey is concluded in Section IX.

Methodology

In this section, the methodology of SA on social media is elaborated, encompassing the entire process from data collection to the implementation of classifiers. Within this pipeline, key stages include data processing, Feature Extraction (FE), and FS. The whole process is given in Figure 2.

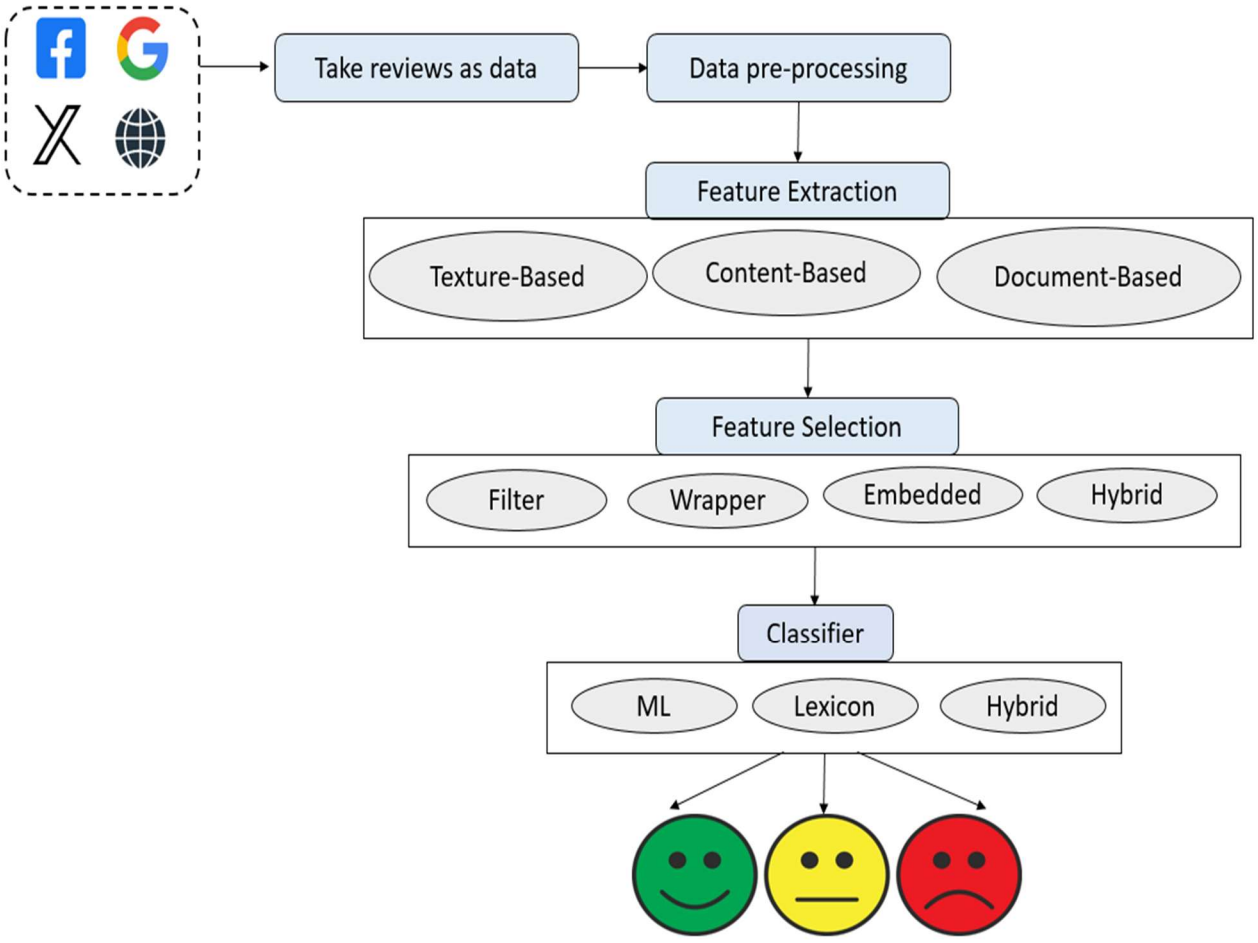

Fig. 2. Work flow of sentiment analysis process

Data Acquisition

To conduct SA effectively, it's essential to have a dataset comprising text data paired with corresponding sentiment labels, such as positive, negative, or neutral. There are numerous publicly accessible datasets available for SA, and a selection of well-known ones is given in Table 1. The table comprises the name of the dataset, its

provided web link, the overall volume of data within the dataset, and the number of distinct classes.

Table 1. Sentiment analysis dataset

Dataset	Ref	Total	Classes
IMDB Movie Reviews	[11]	50,000	2
Yelp Reviews	[12]	500,000	5
Twitter Sentiment Analysis	[13]	1,62,980	3
Airbnb Reviews	[14]	2,686,354	3
Rotten Tomatoes Movie Reviews	[15]	10,48,575	3
Kaggle Sentiment140	[16]	1600000	2
Hotel Reviews	[17]	35,912	5
Financial News Sentiment	[18]	4846	3
Reddit Comments	[19]	37,249	3
MovieLens User Reviews	[20]	25,000,095	Many (0.5 to 5.0)
Twitter Airline Sentiment	[21]	14640	3

Data Processing

The second phase of our research centers on text pre-processing. During this stage, we eliminate any information that is deemed unnecessary from the dataset [22]. Figure 3 below illustrates several pre-processing steps that can be applied to the text data.

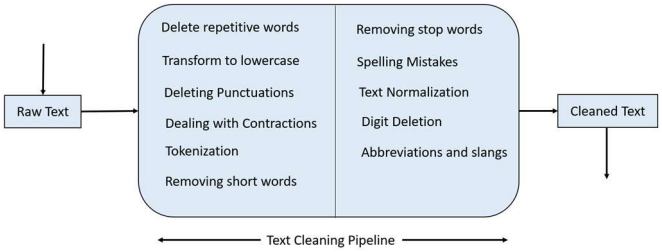

Fig. 3. Data pre-processing techniques

Conversion to lowercase: Converting all characters to lowercase is a crucial pre-processing step, significantly reducing the time required for text processing. While humans easily understand that 'great' and 'Great' are the same, computers treat them as distinct features, necessitating separate processing.

Handling contractions: Contractions, which combine or condense two words into one, are a common language feature. Examples are 'can't' (can + not) and 'wouldn't' (would + not). Many NLP tasks require the expansion of these contractions as a phase of pre-processing.

Tokenization: Tokenization involves breaking down a sequence of data, like textual content, into individual tokens. This can be performed at various levels, including words, sentences, paragraphs, or other meaningful components [23].

Removal of short words: Even after data cleaning, certain meaningless words may persist. To solve this, researchers used a regular expression to exclude terms with two or fewer characters. Because these terms provide no valuable information, they were removed from the dataset.

Elimination of repetitive words: Given the use of Twitter data, it's important to recognize that words with hashtags often repeat frequently and it doesn't give any useful information for training the classifier. Thus, words starting with '@' were removed. For example, mentions of airline names or individuals, while appearing as hashtags, are not valuable for SA and were thus eliminated.

Punctuation removal: Punctuation marks such as commas, exclamation marks, question marks, ellipses, semicolons, colons, and brackets were removed. Some punctuations required separate removal through regular expressions [24].

Digit removal: Digits were excluded from the text since they do not provide essential information for SA, though this exclusion may not apply to every NLP task.

Slangs and Abbreviations: This phase entails fixing internet terminology or acronyms. To improve readability, predefined dictionaries were used to transform abbreviations or slang into their complete forms.

Stop-word removal: Common English words such as 'the,' 'a,' 'an,' and 'in' were removed from the tweet text. These words do not contribute meaningful information to SA.

Spelling correction: Addressing spelling errors is an essential pre-processing step. Users commonly make spelling mistakes, resulting in multiple words with identical root patterns. For example, different people may misspell the term in different ways, leading to new word attributes that have to be analysed, which takes more time.

Text Normalization: Text normalization encompasses techniques like lemmatization [25] and stemming [26] to reduce tokens to their basic forms. Lemmatization produces superior results compared to stemming, despite the additional time required. While aimed to reduce time complexity in SA, the improved quality achieved through lemmatization justifies the extra time invested.

Feature Extraction

In NLP, translating text into numbers is crucial to a computer's understanding of what it reads. Word vectorization, often called word embedding, is a standard FE method [27]. It analyses text by breaking it down phrase by phrase and then word by word. The next step is to carefully construct a feature matrix. Each row of the resulting matrix corresponds to a phrase, and each feature column corresponds to a lexical term. In most cases, the amount of words in a particular phrase or document corresponds to the value in the cell of this feature matrix [28]. There are three types of feature extraction methods based on texture, content and document as shown in figure 4.

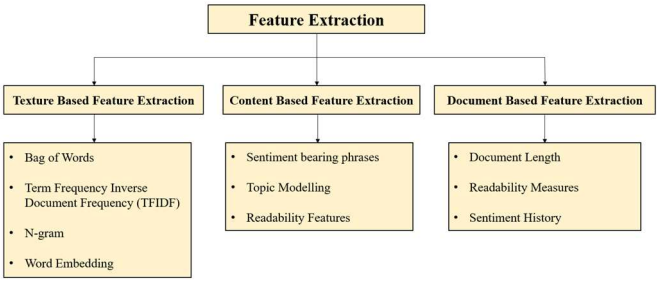

Fig. 4. Types of Feature Extraction approaches

Texture-based FE: When it comes to SA, texture-based FE is indispensable for comprehending the structure and patterns within textual data. It involves a thorough examination of the relationships between words, phrases, and the overall organization of content. These features are invaluable for gaining a nuanced and contextually rich understanding of sentiment. Let's delve into the distinct types of texture-based FE methods utilized in SA:

Bag of Words (BoW): This method simplifies text by stripping it of its sequence and structure, reducing it to what can be described as a "bag" of individual words [29]. The model's focus revolves around the presence or absence of known terms within a document, irrespective of their position. Its objective is to transform each document into a vector that can be efficiently processed by ML models. This FE approach is usually combined with other methods and algorithms to improve its effectiveness.

Term Frequency-Inverse Document Frequency (TF-IDF): The TF-IDF approach is a statistical method that gauges the significance of words within a set of documents [30]. It computes this significance by considering two vital metrics across a collection of texts: the term's frequency within a specific document (TF) and the term's inverse document frequency (IDF). TF-IDF holds substantial value in ML and NLP tasks, particularly those involving text analysis. The formula for computing the TF-IDF score is given below

$$tf - idf(d,t) = tf(t) * idf(d,t)$$

N-grams: N-grams are graphical models that are used to detect patterns in sequential data, like text. An N-gram is a series of "N" objects extracted from a particular text corpus, which can be words, letters, or syllables [31]. N-grams play a crucial role in building language models used in applications like speech recognition and machine translation. These combinations of words or letters create units of meaning and assist in deciphering word context, contributing to NLP tasks.

Word Embedding: The concept revolves around transforming text into numerical vectors [32]. Various approaches are employed. For instance, word2vec, introduced in 2013 by Tomas Mikolov from Google, identifies similarities among words and constructs vectors that represent word features, like word context. This method also makes accurate assumptions about word meanings based on their contexts, enabling associations between words. Glove, developed by researchers at Stanford in 2014, utilizes matrix factorization techniques on word-context matrices to obtain vector representations for words. FastText, introduced by Facebook researchers in 2016, delves even deeper by splitting words into subwords, providing a more comprehensive representation.

Content-based FE: This is an essential component of SA, comprising an in-depth analysis of textual content to extract critical features that provide insights into the conveyed sentiment. These characteristics serve as the foundation for comprehending and categorizing sentiment in documents, communications, reviews, or any other written communication.

Sentiment-Bearing Phrases: The identification of sentiment-bearing phrases is critical in content-based FE. The precise recognition and extraction of specific words, phrases, or expressions embedded within the text that inherently communicate pronounced positive or negative attitudes is required for this technique. These phrases have the ability to drastically impact the overall tone of the text. For example, statements like "over the moon" emanate happiness and ones like "unbearably disappointed" ooze sadness. For SA, this method frequently relies on predetermined lexicons or lists of such words, allowing sentiment analysts to quantify the existence of these sentiment-bearing expressions and evaluate the intensity and orientation of sentiment within the text.

Topic Modeling: Topic modelling is a common and important ML and NLP algorithm. It seeks to extract hidden subjects from lengthy papers [33]. Researchers have shown a strong interest in extracting relevant insights from unstructured short-text sources as social media platforms have grown in popularity. Topic modelling is critical for uncovering underlying themes in tweets and other similar content. The Latent Dirichlet Allocation (LDA) model, for example, is a prime example of a topic modelling technique. LDA utilizes pre-processed materials as input and clusters words inside the documents to form groups of words with similar meanings. Each word in these clusters is assigned a probability based on its relation to the topic of the cluster [34]. These clusters reveal several themes found in the papers.

Readability Features: Another important facet of content-based FE is readability. These characteristics evaluate the complexity and ease of reading a text, which has an indirect impact on the expression of sentiment. The reasoning for this is that complicated texts may transmit sentiment differently than simpler, more straightforward language. Readability methods, such as the Flesch-Kincaid or Gunning Fog Index, assess complexity by taking into account variables such as sentence length and word complexity. A higher Flesch-Kincaid score, for example, suggests more complicated content, which may influence how the audience expresses and comprehends sentiment. Analysts can evaluate the various levels of text difficulty and their implications for sentiment expression by including accessibility factors into SA.

Document-based FE: This method is critical in SA since it aims to capture the overall context and structure of the document under consideration. Document Length, Readability Measures, and Sentiment History are three important forms of document-based FE in sentiment research.

Document Length: Document Length is a simple metric that involves counting the number of words, characters, or phrases in a document. A document's length is a basic yet informative feature. Longer texts frequently allow for more subtle sentiment expressions, hinting at the depth of sentiment portrayed within the language.

Readability Measures: Another essential part of document-based FE is readability measures, which assess the complexity and readability of a document. Algorithms such as the Flesch-Kincaid Grade Level and the Gunning Fog Index use a range of factors to calculate readability scores, including sentence length, word complexity, and syllable count. These scores provide information about the text's complexity, which can indirectly influence how sentiment is portrayed. Text that is more intricate, for example, may express emotion differently than simpler content.

Sentiment History: A chronological approach to Sentiment History involves keeping track of how feelings have evolved over time in relation to a collection of documents. This approach shines brightest when analysing time-stamped content like social media posts, consumer reviews, and the like for shifting sentiment trends. Each document in the series is subjected to SA, and the results are analysed for recurring patterns. You can see the changing sentiment trends in product reviews over time, and here is where Sentiment History comes in handy; it shows you how sentiment changes across a series of documents.

Feature Selection

Several FS techniques are used to reduce irrelevant and redundant characteristics [35]. The goal of FS is to increase the accuracy with which sentiment is classified by identifying and eliminating unnecessary or irrelevant qualities from the feature list [36]. Both statistical and lexicon approaches can be used to choose features. In lexicon-based methods, humans create the features. An initial feature set is often constructed by gathering words that are highly correlated with sentiment. After that, new terms are added by using techniques like synonym identification or consulting web resources to build upon this initial collection. These approaches are effective because they involve a careful selection of features. However, they can be time-consuming and challenging. A well-known example of this approach is the SentiWordNet8 lexicon. In contrast, statistical approaches are entirely automated and widely used for FS. However, they may struggle to distinguish sentiment-carrying features from non-sentiment-related ones. Statistical approaches are typically categorized into four groups [37] which are detailed below and given in Figure 5

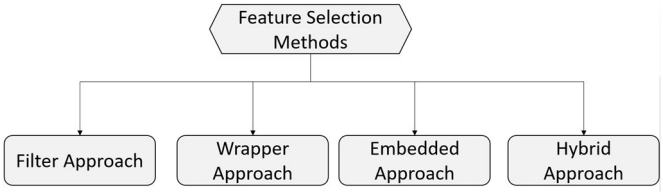

Fig. 5. Types of Feature Selection Approaches

Filter Approach: The most common method of FS does not employ any machine learning techniques, instead selecting features by considering the general attributes of the training data. Features are ranked using statistical measures, and those with the highest scores are chosen. For huge datasets with many characteristics, filter methods are a good option because they require less computing. Chi-square (CHI), Information Gain (IG), Mutual Information (MI), and Document Frequency (DF) are common examples [38].

Wrapper Approach: This method uses ML techniques to assess a subset of characteristics [39]. Wrapper methods are frequently iterative and computationally costly, but they have the benefit of discovering the best-performing feature subset particular to the modelling algorithm used. A wrapper technique integrates ML algorithms with a strategy for generating feature subsets.

Embedded Approach: Using classification algorithms with inherent FS capabilities, FS is easily incorporated into the execution of the modelling algorithm in this manner. Embedded approaches are more efficient than wrapper methods in terms of computational efficiency, but they are fundamentally specialized to the specific learning algorithm being used [40]. Decision tree techniques like C4.5, CART, ID3, and LASSO, are common embedding approaches [41].

Hybrid Approach: This strategy includes features from both the filter and wrapper methods, and it combines many ways in general to retrieve the greatest possible features [42]. Hybrid techniques surpass single techniques with respect to both speed and precision by integrating the best qualities of many approaches. For SA, numerous hybrid FS methods have been developed.

ML Model

Sentiment categorization methods are widely grouped into three approaches: ML, lexicon-based, and hybrid approaches [43]. The ML approach employs well-known supervised and unsupervised algorithms [44]. On the other hand, the lexicon-based method relies on a sentiment lexicon. Dictionary-based (DBA) and corpus-based approaches (CBA) are two more branches of this approach [45]; both employ statistical or semantic techniques to label words as positive or negative depending on their context. The hybrid approach combines ML and lexicon-based techniques for SA [46]. Figure 6 provides an overview of these various methodologies and displays the most commonly used sentiment categorization algorithms.

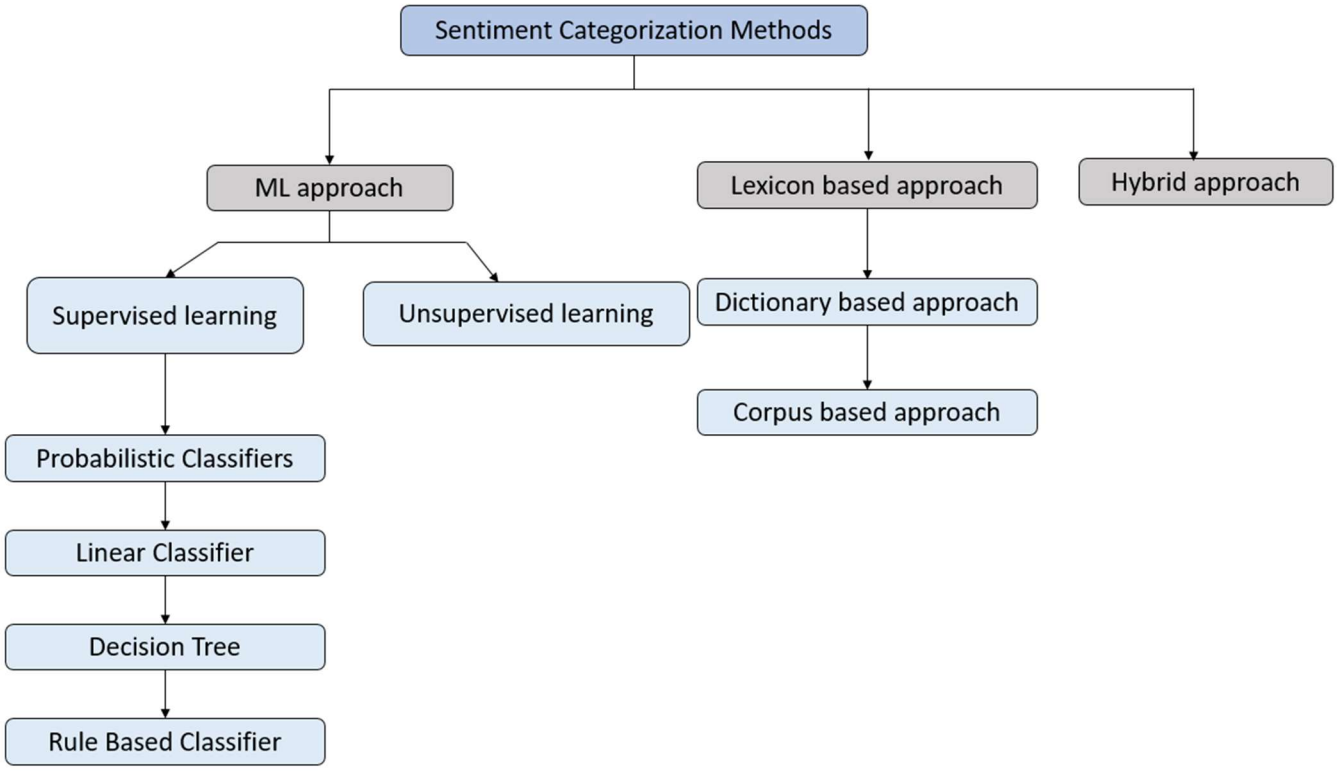

Fig. 6. Types of Classifiers for sentiment analysis

ML approach

The ML method treats SA as a standard text classification problem and employs syntactic and linguistic features to do so [47]. In order to formulate the text classification problem, begin with a training dataset, represented as $D = \{X_1, X_2, \ldots, X_N\}$, in which each record is associated with a specific class label. The classification model makes use of the features present in each record to designate it to one of the class labels. When only a single label is ascribed to an instance, it constitutes a hard classification problem. On the other hand, in the case of soft classification, an instance is assigned a probabilistic distribution of labels.

Supervised learning: These methods in SA depend on labeled training documents. The literature encompasses a variety of supervised classifiers. In the following sections, we will provide a brief overview of some of the most frequently employed classifiers in the field of SA.

Probabilistic classifiers distinguish themselves from linear approaches, which simply designate the most probable class for a given input, whether it be positive or negative [48]. Probabilistic classifiers, on the other hand, estimate a probability distribution across a set of classes, often relying on Bayes' theorem as their foundation. These classifiers employ mixture models for the purpose of classification, treating each class as a component within the mixture. They are frequently referred to as generative classifiers because each element of the mixture acts as a generating model. Probabilistic classifications are simple to design, computationally effective when compared to other techniques, and do not require a large amount of training data. Their ability to correctly classify data, however, could suffer if the data drastically violates the distribution assumptions.

Linear classifier: It utilizes linear or hyperplane decision boundaries for sentiment classification [49]. When working with multiple classes, the word "hyperplane" enters into the scene. The categorization of sentiment is accomplished using a linear predictor, represented as $p = A.X + b$, while A indicates the vector of linear coefficients (weights) and X indicates the document's word frequencies. Predictions are calculated using the dot product of A and X, which is supplemented by the bias term b. When the correct features are used, linear classifiers can achieve extraordinary performance despite their apparent simplicity.

Decision tree classifiers: It create a hierarchical structure within the training data space, using specific attribute conditions to partition the data [50]. These conditions, or predicates, typically revolve around whether certain words are present or absent. The data space is repeatedly divided into smaller segments through a recursive process until the leaf nodes contain a minimum number of records, and this division serves the purpose of classification. Various predicates, based on document similarity and term correlations, can be employed to further segment the data. These techniques for segmentation encompass several approaches. The "Single Attribute Split" method evaluates whether specific words or phrases are present or absent at particular nodes within the tree and leverages this information to perform the split. The "Similarity-based multi-attribute split" strategy involves considering document clusters or sets of frequently occurring words and the similarity of documents to these clusters when making the split. In text classification, the implementation of decision trees typically features minor modifications to standard packages like ID3 and C4.5.

Rule-based classifier: Classifiers that rely on rules to characterize data typically have feature-related conditions written on the left side of their rules in disjunctive normal form. Over on the right, we will find the criteria by which we classify things. Especially in circumstances with sparse data, these criteria focus on the existence of particular phrases rather than their absence. Two of the most common criteria used for rule generation during training are "support," which measures the total number of instances in the training data associated with the rule, and "confidence," which indicates the conditional probability that the right-hand side of the rule is true if the left-hand side is satisfied.

Unsupervised learning: This method presents a challenge similar to supervised learning in that it deals with unannotated or unlabelled datasets for training. This complexity sets it apart from the supervised approach. Unsupervised learning endeavours to reveal concealed patterns and structures in unannotated data without requiring pre-existing training. It is often employed to group or cluster data into categories based solely on their statistical properties. Unsupervised learning methods aim to identify patterns and viewpoints within individual words or phrases and then classify each document based on the presence and frequency of these words or phrases in each document. Common examples of unsupervised learning techniques include clustering algorithms like principal component analysis, k-means, modified k-means, matrix factorization, expectation maximization algorithm, and various others.

Lexicon-based approach

The lexicon-based method for SA hinges on a dedicated lexicon that encompasses a carefully curated assortment of sentiment-related terms. Within this lexicon, each term is assigned a polarity value, where positive words possess values exceeding zero, negative words exhibit values below zero, and any term not present within the lexicon is regarded as neutral [51]. In practical terms, this approach involves performing sentiment classification by searching for sentiment words within a given text or document. After identifying these sentiment

words, weights or tags are assigned to them, and these weighted tags are tallied to identify the overall sentiment. The sentiment lexicon, containing the list of sentiment words along with their polarity values, is crucial to this approach. Preparing such a lexicon can be accomplished through two main methods: the DBA and CBA.

Dictionary-based Approach: This method commences with an initial set of foundational sentiment words that are already associated with known positive and negative orientations [52]. Subsequently, it harnesses linguistic resources like thesauri and corpora, such as WordNet, to detect synonyms and antonyms for each word within the foundational list. The newly unearthed words are then incorporated into the foundational list, and the process continues iteratively. This iterative process concludes when no further new words can be identified. A notable limitation of this approach lies in its incapacity to identify opinion words with domain-specific orientations. For example, if we examine the phrase "the phone speaker is quiet," it conveys a negative sentiment, while "the car is quiet" signifies a positive sentiment.

Corpus-based Approach: Unlike the DBA, the CBA excels at uncovering opinion words that are specific to particular domains and contexts [53]. This approach depends on statistical or syntactic patterns in conjunction with an initial list of opinion words with established polarities to identify fresh sentiment words within an extensive text corpus. In statistical pattern analysis, novel sentiment words are recognized based on their frequency of occurrence within a sizable, annotated corpus. If a word emerges more frequently in positive documents than in negative ones, it is included in the list as a positive word, and conversely for negative occurrences. In other words, if a word is more common in positive documents, it is classified as positive, and if it is more prevalent in negative documents, it is categorized as negative. On the flip side, syntactic pattern analysis aims to detect sentiment words by examining their tendency to appear alongside other words within the corpus. This method operates on the premise that words frequently found together in documents are likely to share the same polarity. If a word does not have an assigned polarity value but is frequently observed alongside another word with a known polarity, it is ascribed the same or opposite polarity based on the connecting word between them, such as the term "AND." For instance, in the sentence "the car is comfortable and spacious," the word "spacious" would be attributed the same polarity as "comfortable" through this process.ble 2. Comparison of past research work on sentiment analysis

Hybrid approach

The hybrid approach to SA seamlessly combines the advantages of both lexicon-based and ML approaches [54]. It integrates the power of lexicon analysis with the versatility of ML methods enabling it to negotiate complexities and identify the sentiment. The goal of harnessing the precision of ML while retaining the robustness inherent in lexicon-based approaches is the fundamental driver for adopting the hybrid approach. This hybrid strategy integrates methods from the preceding approaches in order to conquer their specific drawbacks and maximize their various advantages. To do this, the sentiment ratings generated by lexicon-based research are employed as input features for the sentiment classification. Sentiment lexicons, which are well-known for boosting performance, play a significant role in this approach. It is worth mentioning that only a few models in SA use the hybrid approach, with the bulk using lexicon-based techniques to classify word polarity before integrating it into the SA classifier. An early example of this hybrid strategy is found in the article [55], which merged ML classifiers with dictionaries and HARN's method, a lexicon-based classifier, to categorize documents. Initially, the reviews are identified using two ML classifiers: Nave Bayes and Support Vector Machines (SVM). Following that, they used HARN's technique to determine the sentiment polarity. When compared to HARN's method alone, the hybrid technique displayed a considerable gain in accuracy, achieving roughly 80% to 85% accuracy. Deep learning

methods can also be used with lexicons to do SA jobs.

Previous Research

Table 2 summarizes past work on sentiment analysis using ML. The table explains the research by providing the necessary elements such as data, pre-processing, feature extraction and selection, classifier technique, and accuracy

Reference	Data	Pre-process	Feature Extraction	Feature Selection	Classifier	Accuracy
[56]	IMDb movie reviews	Tokenize, Lemmatization	TF-IDF	-	SVM	89.20%
[57]	IMDb movie reviews	Lemmatization, Punctuations Removal, Stopword Removal,	Hybrid features (TF, TF-IDF with Lexicon feature)	Correlation	Maximum Entropy	83.93%
[58]	Yelp review datasets	Lemmatization, Stopword Removal,	TF-IDF	-	BERT	97.3%
[59]	Yelp review datasets	Tokenize, Lemmatization, Stemming	TF-IDF	Correlation	SVM	98%
[60]	Rotten Tomatoes Movie Reviews	Tokenize, Lemmatization, Stopword Removal, Stemming	Word2Vec	-	Modified Balanced Random Forest	84.15%
[61]	Sentiment140 Twitter dataset	Lemmatization, Stopword Removal, Punctuations	TF-IDF	-	LSTM	82.93%

		Removal, Stemming				
[62]	Twitter US Airlines	Stopword Removal, URL Removal, Special Character Removal	TF-IDF	-	SVM	71.84%
[63]	Twitter US Airlines	Stopword Removal	Word Embedding	-	universal language model fine-tuning-SVM	99.78%

Challenges

In the age of the Internet, an abundance of informal text data is being generated by people. The expansion of social networking platforms has generated various obstacles, such as spelling errors, the development of different slang, and improper usage of grammar. These obstacles make sentiment and emotion analysis more difficult. There are instances where individuals do not convey their emotions in a straightforward manner. Take "I missss u soooo much?" as an example. In this sentence, "miss" is misstyped as "missss," "you" is abbreviated to "u," and "soooo" is employed for added emphasis. Furthermore, the sentence does not clearly indicate whether the person is expressing anger or concern. Therefore, the endeavor of extracting sentiment from real-time data is inherently intricate for various reasons [64].

One prevalent challenge encountered in the realms of emotion recognition and SA pertains to resource scarcity. Certain statistical algorithms necessitate extensive annotated datasets. However, while collecting data may not be arduous, manually labeling large datasets is a time-consuming and somewhat unreliable process [65]. Additionally, the issue of resource availability primarily revolves around the dominance of the English language. As a result, performing SA and identifying emotions in languages other than English, especially regional languages, provides a significant difficulty for scholars. Furthermore, many corpora and lexicons are domain-specific, restricting their usefulness in a variety of scenarios.

Another common issue often observed in posts and conversations on social media like Facebook, Twitter, and Instagram pertains to the usage of web slang. For instance, words such as "OMG" (Oh My God) to express amusement, or "WOW" (Wide-eyed Open Wonder) to convey surprise, are widely used by the younger generation. Conventional lexicons and trained models face significant challenges as the vocabulary of web slang grows.

Furthermore, people frequently communicate their rage or displeasure with sarcasm and irony, rendering it difficult to discern their genuine feelings [66]. Look at the sentence: "This tale is fantastic for getting you to sleep." While the phrase "fantastic" may normally indicate a pleasant attitude, the reviewer thought the plot to be

fairly uninteresting. As a result, detecting sarcasm has grown to be a difficult task in the field of sentiment identification.

Another challenge emerges when attempting to convey more than one feeling in a single phrase. Determining the underlying feelings or sentiments of a text with multiple viewpoints can be a difficult task. Example: "The scenery at this location is so peaceful and tranquil, but the location stinks," where two feelings, "disgust" and "soothing," are conveyed in different ways. Additionally, detecting polarity from comparative sentences proves to be intricate. Consider the two sentences: "Movie 1 is poorer than Movie 2" and "Movie 2 is poorer than Movie 1." In both sentences, the term "poorer" signifies a negative polarity, yet the two sentences hold contrasting meanings [67].

Conclusion

In this survey, the importance of SA in the modern socio-economic landscape is underscored. The ability to understand and effectively employ SA methods across a wide range of input formats is now seen as crucial for the success and longevity of businesses, institutions, and individuals in our data-driven world. SA, or the process of extracting and interpreting sentiment and opinions from textual data, has far-reaching implications. This paper delves into the multifaceted applications of SA and the challenges that come with it, especially in the context of classifying sentiments within single domains and across domains. Accurately gauging public opinion is critical for making sound choices, responding to market trends, and reacting to public input. One of the most important findings from this survey is the understanding that a number of pre-processing methods are required before sentiment categorization can be performed properly. The study also examines feature extraction, recovery of relevant information from text, and the selection of relevant features to improve the accuracy of SA. The survey exhaustively describes the complete SA pipeline, from data collection to classifier implementation. In a world where data is abundant and communication is primarily digital, the capacity to effectively evaluate public sentiment is essential. The concepts and strategies introduced in this paper aim to improve the comprehension of SA and provide valuable support to emerging researchers as they delve into this field.

Reference

1. . Yue, Lin, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. "A survey of sentiment analysis in social media." Knowledge and Information Systems 60 (2019): 617-663.
2. Martínez-Cámara, Eugenio, M. Teresa Martín-Valdivia, L. Alfonso Urena-López, and A. Rturo Montejo-Ráez. "Sentiment analysis in Twitter." Natural language engineering 20, no. 1 (2014): 1-28.
3. Han, Hyun Jeong, Shawn Mankad, Nagesh Gavirneni, and Rohit Verma. "What guests really think of your hotel: Text analytics of online customer reviews." (2016).
4. Zheng, Xiaolin, Shuai Zhu, and Zhangxi Lin. "Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach." Decision Support Systems 56 (2013): 211-222.
5. Macanovic, Ana. "Text mining for social science–The state and the future of computational text analysis in sociology." Social Science Research 108 (2022): 102784.
6. Kennedy, Brendan, Ashwini Ashokkumar, Ryan L. Boyd, and Morteza Dehghani. "Text analysis for psychology: Methods, principles, and practices." (2021).

7. Weber, Charlotte Teresa, and Shaheen Syed. "Interdisciplinary optimism? Sentiment analysis of Twitter data." Royal Society open science 6, no. 7 (2019): 190473.
8. Liu, Bing. Sentiment analysis and opinion mining. Springer Nature, 2022.
9. Li, Jiandun, Pin Lv, Wei Xiao, Liu Yang, and Pengpeng Zhang. "Exploring groups of opinion spam using sentiment analysis guided by nominated topics." Expert Systems with Applications 171 (2021): 114585.
10. Tubishat, Mohammad, Norisma Idris, and Mohammad AM Abushariah. "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges." Information Processing & Management 54, no. 4 (2018): 545-563.
11. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data
12. https://www.kaggle.com/datasets/omkarsabnis/yelp-reviews-dataset
13. Hussein, Sherif. "Twitter Sentiments Dataset." Mendeley Data, V1 (2021).
14. Alsudais, Abdulkareem, and Timm Teubner. "Large-scale sentiment analysis on airbnb reviews from 15 cities." (2019).
15. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/
16. https://www.kaggle.com/datasets/kazanova/sentiment140
17. https://www.kaggle.com/datasets/datafiniti/hotel-reviews
18. Malo, Pekka, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. "Good debt or bad debt: Detecting semantic orientations in economic texts." Journal of the Association for Information Science and Technology 65, no. 4 (2014): 782-796.
19. https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset?select=Reddit_Data.csv
20. Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." Acm transactions on interactive intelligent systems (tiis) 5, no. 4 (2015): 1-19.
21. https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
22. Bao, Yanwei, Changqin Quan, Lijuan Wang, and Fuji Ren. "The role of pre-processing in twitter sentiment analysis." In Intelligent Computing Methodologies: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10, pp. 615-624. Springer International Publishing, 2014.
23. Yogish, Deepa, T. N. Manjunath, and Ravindra S. Hegadi. "Review on natural language processing trends and techniques using NLTK." In Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018, Solapur, India, December 21–22, 2018, Revised Selected Papers, Part III 2, pp. 589-606. Springer Singapore, 2019.
24. Nunberg, Geoffrey. The linguistics of punctuation. No. 18. Center for the Study of Language (CSLI), 1990.
25. Symeonidis, Symeon, Dimitrios Effrosynidis, and Avi Arampatzis. "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis." Expert Systems with Applications 110 (2018): 298-310.

26. Hidayatullah, Ahmad Fathan, C. I. Ratnasari, and S. Wisnugroho. "The influence of stemming on Indonesian tweet sentiment analysis." In Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015), pp. 127-132. 2015.

27. Bakarov, Amir. "A survey of word embeddings evaluation methods." arXiv preprint arXiv:1801.09536 (2018).

28. Ahuja, Ravinder, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. "The impact of features extraction on the sentiment analysis." Procedia Computer Science 152 (2019): 341-348.

29. Kasthuriarachchy, Buddhika H., Kasun De Zoysa, and H. L. Premaratne. "Enhanced bag-of-words model for phrase-level sentiment analysis." In 2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 210-214. IEEE, 2014.

30. Addiga, Akash, and Sikha Bagui. "Sentiment analysis on twitter data using term frequency-inverse document frequency." Journal of computer and communications 10, no. 8 (2022): 117-128.

31. Ojo, O. E., A. Gelbukh, H. Calvo, and O. O. Adebanji. "Performance study of n-grams in the analysis of sentiments." Journal of the Nigerian Society of Physical Sciences (2021): 477-483.

32. Li, Yang, and Tao Yang. "Word embedding for understanding natural language: a survey." Guide to big data applications (2018): 83-104.

33. Rana, Toqir A., Yu-N. Cheah, and Sukumar Letchmunan. "Topic Modeling in Sentiment Analysis: A Systematic Review." Journal of ICT Research & Applications 10, no. 1 (2016).

34. Schofield, Alexandra Kathryn. Text Processing for the Effective Application of Latent Dirichlet Allocation. Cornell University, 2019.

35. Agarwal, Basant, and Namita Mittal. "Optimal feature selection for sentiment analysis." In Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14, pp. 13-24. Springer Berlin Heidelberg, 2013.

36. Ahmad, Siti Rohaidah, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. "A review of feature selection techniques in sentiment analysis." Intelligent data analysis 23, no. 1 (2019): 159-189.

37. Prastyo, Pulung Hendro, Igi Ardiyanto, and Risanuri Hidayat. "A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods." In 2020 6th International Conference on Science and Technology (ICST), vol. 1, pp. 1-6. IEEE, 2020.

38. Harish, B. S., and M. B. Revanasiddappa. "A comprehensive survey on various feature selection methods to categorize text documents." International Journal of Computer Applications 164, no. 8 (2017): 1-7.

39. Suchetha, N. K., Anupama Nikhil, and P. Hrudya. "Comparing the wrapper feature selection evaluators on twitter sentiment classification." In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1-6. IEEE, 2019.

40. Rokon, Md Omar Faruk, Pei Yan, Risul Islam, and Michalis Faloutsos. "Repo2vec: A comprehensive embedding approach for determining repository similarity." In 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 355-365. IEEE, 2021.

41. Liu, Haoyue, MengChu Zhou, and Qing Liu. "An embedded feature selection method for imbalanced data

classification." IEEE/CAA Journal of Automatica Sinica 6, no. 3 (2019): 703-715.

42. Ansari, Gunjan, Tanvir Ahmad, and Mohammad Najmud Doja. "Hybrid filter–wrapper feature selection method for sentiment classification." Arabian Journal for Science and Engineering 44 (2019): 9191-9208.
43. Sham, Nabila Mohamad, and Azlinah Mohamed. "Climate change sentiment analysis using lexicon, machine learning and hybrid approaches." Sustainability 14, no. 8 (2022): 4723.
44. Hiran, Kamal Kant, Ritesh Kumar Jain, Kamlesh Lakhwani, and Ruchi Doshi. Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition). BPB Publications, 2021.
45. Sadia, Azeema, Fariha Khan, and Fatima Bashir. "An overview of lexicon-based approach for sentiment analysis." In 2018 3rd International Electrical Engineering Conference (IEEC 2018), pp. 1-6. 2018.
46. Ahmad, Munir, Shabib Aftab, Iftikhar Ali, and N. J. I. J. M. S. E. Hameed. "Hybrid tools and techniques for sentiment analysis: a review." Int. J. Multidiscip. Sci. Eng 8, no. 3 (2017): 29-33.
47. Altınel, Berna, and Murat Can Ganiz. "Semantic text classification: A survey of past and recent advances." Information Processing & Management 54, no. 6 (2018): 1129-1153.
48. Chen, Huanhuan, Peter Tino, and Xin Yao. "Probabilistic classification vector machines." IEEE Transactions on Neural Networks 20, no. 6 (2009): 901-914.
49. Khairnar, Jayashri, and Mayura Kinikar. "Machine learning algorithms for opinion mining and sentiment classification." International Journal of Scientific and Research Publications 3, no. 6 (2013): 1-6.
50. Vens, Celine, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. "Decision trees for hierarchical multi-label classification." Machine learning 73 (2008): 185-214.
51. Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. "Lexicon-based methods for sentiment analysis." Computational linguistics 37, no. 2 (2011): 267-307.
52. Osman, Aida, and Said Ahmad. "Current trends and research directions in the dictionary-based approach for sentiment lexicon generation: a survey." J Theor Appl Inf Technol 97 (2019): 22.
53. Rice, Douglas R., and Christopher Zorn. "Corpus-based dictionaries for sentiment analysis of specialized vocabularies." Political Science Research and Methods 9, no. 1 (2021): 20-35.
54. Gadri, Said, Safia Chabira, Sara Ould Mehieddine, and Khadidja Herizi. "Sentiment analysis: developing an efficient model based on machine learning and deep learning approaches." In Intelligent Computing & Optimization: Proceedings of the 4th International Conference on Intelligent Computing and Optimization 2021 (ICO2021) 3, pp. 237-247. Springer International Publishing, 2022.
55. Devi, DV Nagarjuna, Thatiparti Venkata Rajini Kanth, Kakollu Mounika, and Nambhatla Sowjanya Swathi. "Assay: Hybrid approach for sentiment analysis." In Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 1, pp. 309-318. Springer Singapore, 2019.
56. Ubaid Mohamed Dahir, Faisal Kevin Alkindy, "Utilizing Machine Learning for Sentiment Analysis of IMDB Movie Review Data," International Journal of Engineering Trends and Technology, vol. 71, no. 5, pp. 18-26, 2023.

57. Harish, B. S., Keerthi Kumar, and H. K. Darshan. "Sentiment analysis on IMDb movie reviews using hybrid feature extraction method." (2019).

58. Areshey, Ali, and Hassan Mathkour. "Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model." Sensors 23, no. 11 (2023): 5232.

59. Lavanya, B. N., P. Deepa Shenoy, and K. R. Venugopal. "Sentiment Analysis of Social Media Reviews using Machine Learning and Word Embedding Techniques." In 2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON), pp. 01-05. IEEE, 2023.

60. Nugraha, Mohamad Rizki, Mahendra Dwifebri Purbolaksono, and Widi Astuti. "Sentiment Analysis on Movie Review from Rotten Tomatoes Using Modified Balanced Random Forest Method and Word2Vec." Building of Informatics, Technology and Science (BITS) 5, no. 1 (2023): 153-161.

61. Singh, Harbhajan, and Vijay Dhir. "HYBRID MODEL FOR SENTIMENT ANALYSIS OF TWITTER DATA." Journal of Data Acquisition and Processing 38, no. 2 (2023): 2841.

62. Semary, Noura A., Khalid Amin, and Mohamed Adel Hammad. "Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study." IJCI. International Journal of Computers and Information 10, no. 3 (2023): 52-57.

63. AlBadani, Barakat, Ronghua Shi, and Jian Dong. "A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM." Applied System Innovation 5, no. 1 (2022): 13.

64. Batbaatar, Erdenebileg, Meijing Li, and Keun Ho Ryu. "Semantic-emotion neural network for emotion recognition from text." IEEE access 7 (2019): 111866-111878.

65. Balahur, Alexandra, and Marco Turchi. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis." Computer Speech & Language 28, no. 1 (2014): 56-75.

66. Ghanbari-Adivi, Fereshteh, and Mohammad Mosleh. "Text emotion detection in social networks using a novel ensemble classifier based on Parzen Tree Estimator (TPE)." Neural Computing and Applications 31, no. 12 (2019): 8971-8983.

67. Shelke, Nilesh M. "Approaches of emotion detection from text." International Journal of Computer Science and Information Technology 2, no. 2 (2014): 123-128.