

# An Extensive Review of Sentiment Analysis Leveraging Artificial Intelligence and Machine Learning

**Mrs. Shweta Sinha**

Assistant professor, Department of Computer Science, National Post Graduate College

**Pranati Kaushik**

Student, scholar, Department of Computer Science, National Post Graduate College

## Abstract-

Because social networking sites are growing so quickly in the Internet age, they are now a vital tool for sharing emotions with people everywhere. Many people convey their opinions or thoughts using text, images, music, and video. On the other hand, text communication through Web-based networking media might be a little overwhelming. Social media platforms are the source of vast amounts of unstructured data generated on the Internet every second. To understand human psychology, data must be analysed as quickly as it is generated. Sentiment analysis, which identifies polarity in texts, can help with this. It evaluates the author's attitude toward a thing, organization, person, or place—whether it be favourable, negative, or neutral.

Organizations can obtain real-time insights on customer sentiment, experience, and brand reputation with the use of sentiment analysis technologies. These technologies typically evaluate online sources like emails, blogs, reviews, customer service tickets, news stories, survey results, case studies, web chats, tweets, forums, and comments using text analytics. Whether the consumer is using positive, negative, or neutral language, algorithms are employed to construct rule-based, automatic, or hybrid methods of scoring.

Sentiment analysis may not be sufficient in certain applications; in such cases, emotion detection—which accurately assesses a person's emotional and mental state—is necessary. Understanding of sentiment analysis levels, different emotion models, and the procedure for sentiment analysis and emotion detection from text are all provided by this review work. This paper concludes by discussing the difficulties encountered in sentiment and emotion analysis.

## 1. Introduction

Natural language processing (NLP) has two components: human language generation and human language understanding. However, because genuine language contains ambiguity, the former is more challenging. But because genuine language contains ambiguity, the former is more difficult. NLP is used in speech recognition, machine translation, question answering, speech synthesis, and other applications. The two most important aspects of natural language processing are emotion recognition and sentiment analysis. These two names are not the same, despite the fact that they are commonly used synonymously. Data can be evaluated to determine whether it is favourable, negative, or neutral using sentiment analysis.

On the other hand, emotion detection is the process of recognizing different kinds of human emotions, like anger, happiness, or depression. There are instances when terms like "emotion detection," "affective computing," "emotion analysis," and "emotion identification" are used interchangeably. Social media has become a popular way for people to express their emotions since Internet connectivity has improved. People freely communicate their thoughts, disagreements, and ideas on a wide range of subjects via social media. Additionally, a large number of customers post reviews and comments on a range of goods and services on different e-commerce websites. Customers' evaluations

and ratings across a variety of channels incentivize suppliers and service providers to improve their current products, services, or systems.

An essential tool for businesses to comprehend how consumers view and interact with their brands and products is sentiment analysis. Customer input is becoming more and more dispersed online via several disconnected channels, such social media posts and product reviews on Amazon.

Usually, businesses lack the time and resources to search the internet for information on their products, services, and brand, read through it all, and evaluate it. Instead, they automate this procedure and deliver real-time feedback using sentiment analysis techniques. Businesses use these comments to enhance their offerings in terms of goods, services, and clientele.

## 2. Types of sentiment analysis

Sentiment analysis systems can be classified into multiple categories:

Sentiment indicators are divided into more specific categories, such as extremely positive and extremely negative, using **fine-grained sentiment analysis**. This method is comparable to the one-to-five-star opinion rating system. For this reason, this method works well for rating customer satisfaction surveys.

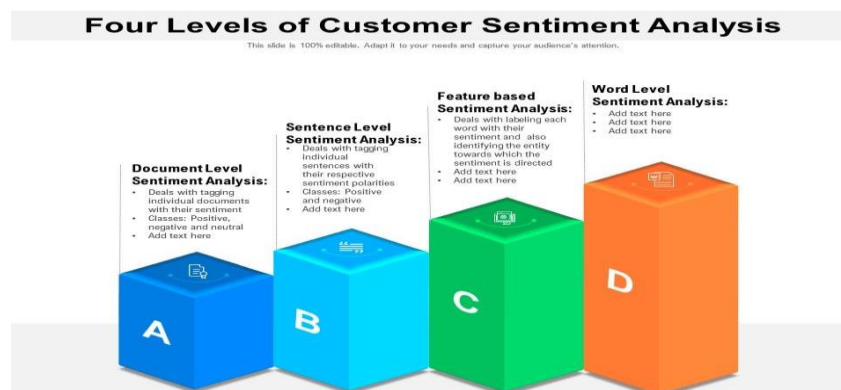
Rather than identifying positive and negative, emotion detection analysis does so. Happiness, dissatisfaction, shock, anger, and grief are a few examples.

**Intent-based analysis** distinguishes between a text's opinions and motivations. For instance, posting a dissatisfied online comment about needing to change a battery could be intended to trigger a customer support contact to address the problem.

**Aspect-based analysis** focuses on the specific component that has been mentioned positively or negatively. For example, a customer may leave a review of a product stating that the battery life was too short. The sentiment analysis system will recognize that the bad sentiment is not about the product as a whole, but rather about the battery life.

## 3. Levels of Sentiment Analysis

Numerous levels of sentiment analysis have been studied, including the document, sentence, phrase, and aspect levels. Sentiment analysis is displayed in Fig. 1 at every level, including the document sentence, and phrase levels.



### 3.1 Document level sentiment analysis

Document-level: Sentiment analysis at the document level is done on the entire document and it is assigned a single polarity. It can be applied to categorize literature pages or chapters as neutral, negative, or positive. At this stage, the material can be classified using supervised and unsupervised learning techniques. The two most important problems in document-level sentiment analysis are cross-domain and cross-language sentiment analysis. It has been demonstrated that domain-specific sentiment analysis is extremely sensitive to domain changes and can attain exceptional accuracy. The feature vector for these tasks is a restricted and domain-specific set of words.

### 3.2 Sentence level sentiment analysis

Sentence level: Each sentence is examined at this level of analysis, and a corresponding polarity is found. When a document is associated with a diverse range of sentiments, this becomes quite helpful. Subjective classification is linked to this level of classification. With more training data and processing power, each sentence's polarity will be assessed individually utilizing the same techniques as the document level. To determine the overall sentiment of the document, the polarity of each sentence can be used separately or together. Sometimes, document-level sentiment analysis is not enough for a given application. Finding subjective sentences has been the focus of earlier sentence-level analytic efforts.

However, more complex jobs, such as working with conditional sentences or ambiguous assertions. Sentence-level sentiment analysis is crucial in these situations.

### 3.3 Phrase level sentiment analysis

Phrase level: Sentiment analysis will also be carried out, which will involve classifying and mining opinion words at the phrase level. There could be one aspect or several aspects in each phrase. Reviews of numerous lines of products could benefit from this; in this case, a single aspect is represented in a single phrase. It has recently become a popular subject among researchers. Sentence-level analysis is more advantageous because a document contains both positive and negative statements, whereas document-level analysis focused on classifying the entire document as subjective, either positively or negatively. The polarity of a word, which is the fundamental building block of language, is closely linked to the subjectivity of the sentence or text in which it appears.

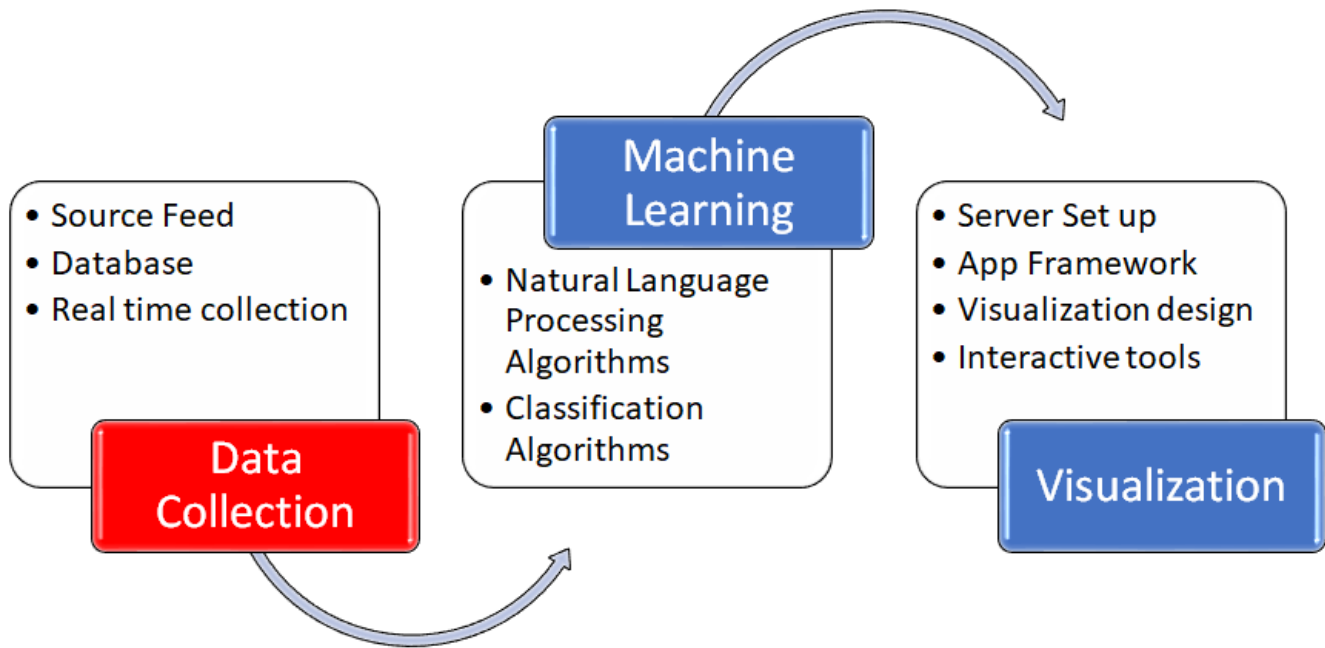
Furthermore, the term selected for expression embodies the demographic attributes of people, including their age and gender, as well as their desire, social status, personality, and other psychological and social traits. Thus, word acts as the basis for sentiment analysis in text.

### 3.4 Aspect level sentiment analysis

Aspect level: An aspect-level sentiment analysis is carried out. Sentiment analysis at the aspect level is necessary since a sentence can have several aspects. After giving each element in the sentence primary consideration and giving it polarity, the overall sentiment for the sentence is determined.

## 4. Gathering information and choosing features

Web scraping, social media, news channels, e-commerce websites, forums, weblogs, and a few other websites depicted in Fig. 2 are some of the ways that data can be gathered from the internet. The initial step in the sentiment analysis process is data collection. Text data can be merged with different types of data, such as audio, video, location, etc., depending on task sentiment analysis of findings. A few crucial places to get data are:



**Figure-2. The process of data collection which follows while performing sentiment analysis.**

#### 4.1 Information Gathering

**Social media:** Information obtained through social media networks is referred to as social data. It illustrates how users access, post, and exchange information about the product. Social media is a dynamic data source used in academic research on individual, group, and behavioral behavior. It describes web-based or mobile applications on the Internet that let users exchange, create, and consume user-generated content.

**Forums:** Users can use message boards to text for help, debate a range of subjects, and exchange thoughts and opinions. Because user-generated content on forums is dynamic, it's a fascinating source for sentiment research. Additionally, by using forums as a source, researchers can perform sentiment analysis on a particular domain.

**Weblog:** A brief weblog is made up of paragraphs that share information, opinions, links, or personal journal entries. Collectively referred to as postings, they are organized chronologically in the manner of a research article, with the most recent entry showing up first. A useful tool for conducting sentiment analysis on a range of things is a blog.

**Websites for electronic commerce:** Websites for electronic commerce on which consumers can rate and comment on a specific company or group. In this case, a descriptive analysis of the different airline service classifications was done

by websites that do not specifically review sites and include millions of reviews, such as e-commerce sites that feature product reviews or expert reviews.

## 4.2 Choosing Features

It's crucial to keep in mind that finding pertinent features in a dataset is the initial step in creating a classification model. As a result, during model training, a review can be translated into English and added to the feature vector. The technique is known as a "Uni-gram" when it considers a single word, a "Bi-gram" when it considers two words, and a "Tri-gram" when it considers three words. The combination of unigram and bigram helpful for analysis of the context feature which is helpful for getting the most accurate results.

Characteristics that prioritize the use of language over a methodological basis are called **pragmatic**. In linguistics and allied sciences, pragmatics is the study of the relationship between context and perception. The study of pragmatics includes investigating phenomena like implicature, speech acts, relevance, and dialogues.

**Emojis** are facial expressions that are used to represent emotions in sentiment analysis. A vast range of human emotions are represented by different emoticons. Emoticons facilitate sentiment analysis by helping to express a writer's tone of voice in a sentence. Replace the emoticons with their meaning: Reviewers express a variety of feelings, such as fury, despair, and delight. Positive and negative sentiment feelings are the two categories into which emotions are divided. Negative emojis are made up of negative emotions like despair, depression, and rage, and positive emoticons are made up of positive emotions like love, joy, and happiness.

**Exclamation points and other punctuation** are used to emphasize the strength of a statement, whether it be good or negative. Other punctuation symbols include the question mark and the apostrophe.

## 4.3 Feature Deletion

A crucial step in sentiment classification is feature extraction, which entails taking important information out of the text data and has an immediate effect on the model's performance. The method is to extract meaningful data that summarizes the key ideas in the text. It is difficult to extract features from the text, hence works contain additional features. Punctuations are often eliminated from text after it has been reduced during the pre-processing phase; nevertheless, they can be used to extract features. The hashtags and emoticons that are frequently employed for feature extractions are given below.

**Terms' regularity** It's one of the easiest ways to represent characteristics that are more commonly utilized for information retrieval in a variety of NLP applications, such as sentiment analysis. The term count of a single word, or unigram, or a combination of two to three words, expressing features in bi- and tri-grams, is taken into consideration. The existence of the term assigns a value of either 0 or 1. Term frequency is the number representing how many times a term appears in the supplied material. For improved results, TF-IDF can be utilized as a weighted system to determine each token's relevance inside the provided document.

**Tags for parts of speech** Grammatical tagging is another name for the practice of labelling a word in a text (corpus) according to its definition and context. Nouns, verbs, pronouns, adverbs, adjectives, and prepositions are the different categories for tokens. For example, the phrase "This phone is amazing" may have the following tags: This is a phone noun, determiner, verb, and magnificent adjective. An adjective is more frequently employed in sentiment mining because it expresses the opinion's feeling. For this task, POS taggers, which are available in NLTK or Spacy, may be utilized.

Neutral sentiment is sometimes included as well, but because neutral evaluations are ambiguous and lacking in detail, they are usually disregarded in sentiment analysis assignments. Numerous studies have attempted to strengthen neutrality by drawing a line between favourable and unfavourable assessments in order to enhance model performance. They collect sentiment from different corpora using a variety of sentiment analysis techniques, then eliminate neutral assessments by consensus—that is, by using several models based on weighted aggregation. Ultimately, they contrasted the classification performance of individual and combined models. Additional contributions pertain to opinion analysis, specifically focusing on multi-level fine-scaled sentiment detection that manages ambivalence

**Words in a Bag (BOW)** One of the easiest methods for extracting text features is BOW. The word occurrences in a document will be described using BOW. Bag is a representation of the word vocabulary that each sentence's vector is created from. The primary issue with this paradigm is that it ignores the text's syntactic meaning. Take two statements, for example:  $s_1 = \text{"The cuisine was good,"}$   $s_2 = \text{"The staff behaviour was bad."}$  The vocabulary is made up of two sentences:  $v_1 = [1 \ 1 \ 1 \ 0 \ 0 \ 1]$  and  $v_2 = [1 \ 0 \ 1 \ 1 \ 0]$ , where  $v = \{\text{'the', 'cuisine', 'was', 'staff behaviour', 'bad', and 'good'}\}$ . The vector has a length of six. The BOW method performs better when its performance is assessed using TF-IDF.

CUSTOMER REVIEW	SENTIMENT
One of the best Indian restaurant in downtown Vancouver.	Positive
I wouldn't recommend this place to anyone. The food is terrible.	Negative

**Figure-3. The process of drawing conclusions by extracting text features.**

## 5. Need of Sentiment Analysis

Sentiment research is very important since it enables organizations to comprehend how customers feel about their brands. Organizations can make educated judgments by automatically categorizing the emotions underlying social media interactions, reviews, and more. Sentiment analysis is the term used to describe the techniques and approaches used by businesses to look at data on the opinions of their clientele regarding a particular service or good. How do I determine the Polarity? shows if a feeling is positive or bad. Subject: What topic is being discussed? Who is the opinion bearer here? something or someone who expresses the feeling.

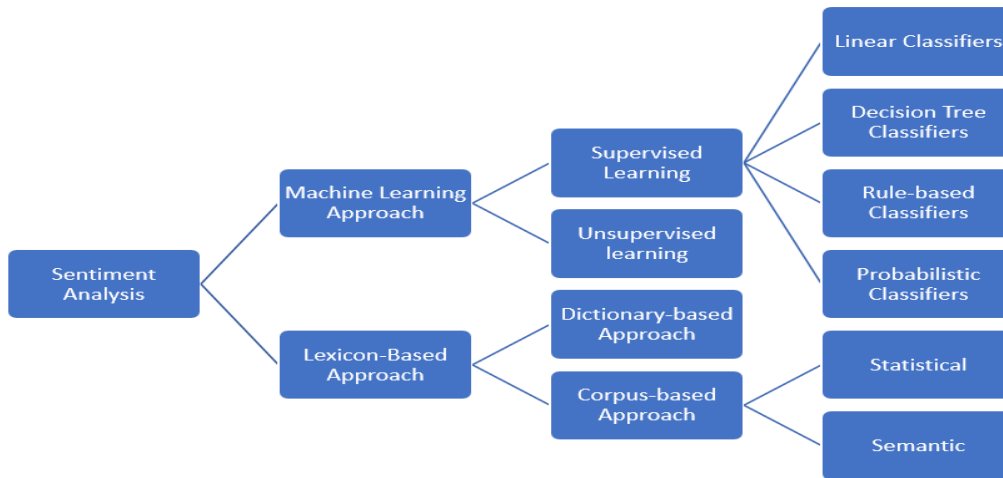
Sentiment analysis is an automated approach that examines natural language speech, extracts salient assertions or viewpoints, and categorizes them based on the emotional disposition of the speakers.

- Sentiment research for business purposes has raised customer satisfaction through improved products, in-the-moment issue identification, and unique market positioning.
- Sentiment analysis of customer satisfaction: Using natural language remarks, the customer discusses his experience with a product and expresses his opinion and attitude about it. This gives us vital information about the customer's level of satisfaction and, if needed, how we can make the product better.



- Recognize issues in real time and take action: A customer can instantly express his displeasure to the world on social media.

## 6. The Sentiment Analysis Methodology



**Figure-3.**

The Lexicon Based Approach, Machine Learning Approach, and Hybrid Approach are the three most used methods for Sentiment Analysis. Additionally, scientists are always looking for more efficient ways to complete the task at hand that need less computing power and improve accuracy.

### 6.1 The Lexicon based approach

Sentiment analysis at the feature and sentence levels is very achievable using the lexicon-based method. It might be referred to as an unsupervised technique since no training data is needed. Nevertheless, as words can have multiple meanings and interpretations, a term that is good in one domain may be negative in another. This leads to the principal drawback of this technique: domain dependence. For example, the word "small" in the sentences "The TV screen is too small" and "This camera is extremely small" has a negative connotation because people typically prefer large screens, but in the second sentence it has a positive connotation because being small will make the camera easier to carry.

This problem can be solved by creating a sentiment lexicon tailored to a particular domain or by modifying an already-existing language.

One benefit of the lexicon-based technique is that it doesn't require any training data, and some experts even refer to it as an unsupervised approach (Yan-Yan et al. 2010). The primary drawback of the lexicon-based method is its strong domain specialization, which prevents terms from one domain from being utilized in another. Take the word "huge," for example. Depending on the context, it can have a positive or negative meaning. Words like "there was a huge lag in network" and "the queue for the movie was huge" can be interpreted positively or negatively.

## 6.2 The Machine learning approach

Sentiment analysis can be done with machine learning algorithms. Sentiment analysis is the process of employing text analysis, computational linguistics, natural language processing, and other methods to identify and quantify the sentiment of text or voice. Sentiment classification is a standard text classification problem that is addressed by the machine learning technique using syntactic and/or linguistic characteristics. The features of the underlying record are linked to one of the class labels by the categorization model. Subsequently, the model is employed to forecast a class label for a specific instance of an unidentified class. We face a challenging classification problem when an instance is given only one label. The term "soft classification issue" refers to the probability value of labels assigned to an instance. Systems can learn new skills without having to be specifically programmed to do so thanks to machine learning. Algorithms for sentiment analysis can be trained to read more than just definitions; they can be taught to understand sarcasm, context, and improper word usage.

## 6.3 The Hybrid approach

Machine learning and lexicon-based techniques are combined in a hybrid method. The word "hybrid" describes sentiment analysis methods that combine lexicon-based and machine learning techniques. Most systems use the hybrid technique, which mixes the two and is very popular, with sentiment lexicons playing a big part. Sentiment analysis employs a hybrid methodology that combines knowledge-based and statistical techniques to identify polarity. Numerous studies suggested combining SVM with two feature selection methods based on the multi-verse optimizer and Relief algorithms to create a hybrid machine learning methodology. tasks for sentiment analysis that use a hybrid machine learning approach that combines RF and SVM.

According to the results, their model performs better than the majority of models yet up to 96% fewer characteristics are needed overall. They also discussed the capabilities of hybrid models and came to the conclusion that, with the right design and careful hyperparameter selection, hybrid models may perform better than any other model. In every other statistic and comparison, the hybrid model fared better than the model alone. They came to the conclusion that even though their hybrid model outperforms individual models, there are still plenty of research opportunities to fine-tune and train the hybrid model to increase its performance.

## 7. Conclusion

The approaches related to sentiment analysis were covered in this article. This work's main goal is to thoroughly examine categorization techniques and weigh the benefits and drawbacks of each for sentiment analysis. We started by talking about the various levels of sentiment analysis and then quickly went over the steps that must be taken, such gathering data and choosing features. Next, different approaches to sentiment categorization systems were categorized and contrasted based on their benefits and drawbacks. In this field, supervised machine learning techniques are frequently the most widely used because of their ease of use and high accuracy.

NB and SVM classification algorithms are frequently used as benchmarks to evaluate newly suggested methods to. Following a discussion of a few of the most popular application domains, the survey looks at the importance and ramifications of sentiment analysis issues in sentiment assessment. The comparison looks into the connection between sentiment reviews' structure and the challenges of sentiment analysis. Domain dependence is shown in this comparison, which is crucial for figuring out sentiment problems. Future research will involve adding new discoveries on a regular



basis to the comparison area. The ensuing difficulties demonstrate that sentiment analysis is a largely unexplored field of study.

### References:

- [1] Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* **55**, 5731–5780 (2022).
  - [2] Qianwen Ariel Xu, Victor Chang, Chrisina Jayne. A systematic review of social media-based sentiment analysis: Emerging trends and challenges, *Decision Analytics Journal*, Volume 3, 2022, 100073, ISSN 2772-6622.
  - [3] A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia *Appl. Sci.* 2023, *13*(7), 4550; <https://doi.org/10.3390/app13074550>
  - [4] A Guide to Text Classification and Sentiment Analysis by Abhijit Roy (July11,2020).
  - [5] A comprehensive survey on sentiment analysis: Approaches, challenges and trends LAROSERI Laboratory, Computer Science Department, University of Chouaib Doukkali, Faculty of Sciences, El Jadida, Morocco
- Received 1 July 2020, Revised 25 March 2021, Accepted 10 May 2021, Available online 14 May 2021, Version of Record 18 May 2021. Volume 226, 17 August 2021, 107134.