

Sentiment Analysis: A Comprehensive Review of NLP Techniques on Social Media Data

Ritish Kumar Sharma^{1,*}, Ankit Kumar², Simran Kumari², Tuba Fatma², Pavitr Jain²,
Sujeet Kumar³

Abstract

The growth of social media platforms has created a large volume of written content generated by users via interactions. The data consists of individuals' opinions and feelings expressed about products and services and political and social issues. Manual methods have made it very difficult to process the massive amount of available data. Therefore, users are implementing sentiment analysis—a natural language processing (NLP) tool that allows individuals to understand communications from the public on the internet via analysis. This paper reviews prior research studies on how sentiment analysis uses various methods derived from NLP to analyze content found on social media. Some of these methods include word lists and definitions, machine learning methods, as well as various Deep Learning Techniques. The research paper discusses the various data preparation techniques and the data extraction process, as well as the types of datasets utilized by the author(s) in their research and the evaluation methods of their research. The research paper also assesses some of the challenges associated with using social media data for sentiment analysis (i.e., detecting sarcasm, multilingual data, and lack of data). That said, the research paper also provides an assessment of the strengths and weaknesses of each of the different sentiment analysis techniques being evaluated.

Keywords: Social media, natural language processing (NLP), sentiment analysis, opinion mining, machine learning

INTRODUCTION

The rapid growth of digital technology and the boom in the use of social media have changed the way people talk, connect, and share their opinions and feelings. Social media has become widespread, where people talk about everything from movies and social issues to political opinions. All that talking adds up. Social media is a platform for friends to meet, but it is now a huge database that holds people's current thoughts and feelings on a wide range of subjects. For researchers and businesses, all that writing

*Author for Correspondence

Ritish Kumar Sharma
E-mail: sharmaritish99@gmail.com

^{1,2}Student, Department of Computer Science and Engineering,
Greater Noida Institute of Technology, Greater Noida, Uttar
Pradesh, India

³Faculty/Academician, Department of Computer Science and
Engineering, Greater Noida Institute of Technology, Greater
Noida, Uttar Pradesh, India

Received Date: June 24, 2026

Accepted Date: August 4, 2026

Published Date: September 10, 2026

Citation: Ritish Kumar Sharma, Ankit Kumar, Simran Kumari, Tuba Fatma, Pavitr Jain, Sujeet Kumar. Sentiment Analysis: A Comprehensive Review of NLP Techniques on Social Media Data. International Journal of Algorithms Design and Analysis Review. 2026; 4(2): 35–42p.

is worth a lot. The field of sentiment analysis, also called opinion mining, has grown into a major field of study during the past few years. People use sentiment analysis to discover and categorize public opinions and emotional states that people express through their writing [1]. This field belongs to the domain of natural language processing (NLP). The primary function of sentiment analysis is to determine whether people write positive, negative, or neutral content. The system enables users to identify specific emotional states, including happiness, anger, sadness, and surprise. Users employ sentiment analysis to determine their emotional state through text input. The study of media content through sentiment analysis provides valuable insights that businesses and governments

find useful for understanding public opinion. Sentiment analysis is used in daily activities because it functions with NLP to help identify users' emotional states [2]. Our understanding of sentiments has changed due to the increase in social media and advancements in machine learning (ML) and AI. Originally, when people were trying to determine whether words were happy, sad, or angry, they used basic techniques to compare the word to a list of other emotional words. This worked reasonably well for basic cases, but those methods completely missed the context. They did not recognize sarcasm, jokes, or phrases that changed meaning, such as using "not bad" to mean "pretty good." Therefore, researchers have evolved. They began creating smarter systems using machine learning. These new tools perform a much better job of understanding what people really mean.

Sentiment analysis is now more precise and reliable. Machine learning has significantly improved sentiment analysis systems that use multiple algorithms to determine the meaning behind written text. The system uses Bayesian support vector machines, decision trees, and logistic regression to classify user statements. The development of these models depends on these algorithms [3]. The system uses labeled training datasets to identify patterns, which are then used for prediction tasks. The new technology makes better predictions because it understands the data better than traditional systems, which depend on fixed regulations. The effectiveness and performance of machine learning models depend on the expertise of their developers and the available training data [3]. Machine learning models operate effectively when developers create high-quality feature sets and use appropriate data. Machine learning models require high-quality data, which should include notes for effective learning. Sentiment analysis has improved significantly because advancements in deep learning technology have gotten much better. Deep learning technologies, such as convolutional neural networks, recurrent neural networks, and long short-term memory, are very good at handling large amounts of text data. They get better at it. They can independently figure out things about language on their own. This is because of the ways of building deep learning models, such as the one called bidirectional encoder representations from transformers, or bidirectional encoder representations from transformers (BERT) for short. These are new models like BERT have made it easier for computers to understand what people are saying and what they mean by it [4].

Therefore, computers can figure out how someone feels about something when they read what that person has written. Sentiment analysis, deep learning technology, and BERT are all improving continuously. Performing sentiment analysis on media is difficult. Researchers must deal with many problems. This is because people do not write the way we want them to write. They use slang, abbreviations, and emojis. They make grammatical mistakes. Sentiment analysis of the media must handle all these factors. There is also the problem of sarcasm and irony. Sarcasm and irony completely change the meaning of a post on social media. People use sarcasm and irony intentionally. They want to play with how their mood is understood by others. People also switch languages in their posts on social media. Sometimes, they even mix languages in a post on social media. This makes it difficult for NLP tools to function. Traditional NLP tools are not designed to handle media conversations. Social media conversations are unpredictable, and traditional NLP tools cannot cope with them. This growing amount of content on social media is generating new problems in data processing, data storage, and real-time data analysis. The system requires effective algorithms and scalable computational systems because it must process millions of posts that produce a huge amount of textual content every minute. Researchers are focusing on the development of hybrid systems to improve the accuracy of sentiment detection by integrating lexicon-based, machine learning, and deep learning systems. The remaining sections of this paper provide an overview of previously published articles and summarize the key contributions made in the academic literature regarding sentiment analysis. Next, the paper will explore different NLP methodologies, compare and evaluate available datasets, and compare results obtained using existing methods [4]. In conclusion, this paper discusses the challenges associated with conducting sentiment analysis on social media content and suggests possible future research directions.

LITERATURE REVIEW

Currently, people are consuming a higher number of media. Existing technology can pull textual data that expresses opinions and emotional responses toward many possible topics. Researchers have focused on how individuals convey their emotions through online writing activities. Several specialists have been called to study this discipline. Researchers should build several approaches. This section provides an overview of essential research in the field while presenting the main advancements achieved in social media sentiment analysis methods.

The first studies conducted in sentiment analysis used lexicon-based methods that created dictionaries containing important sentiment terms and their associated polarity values [5]. The initial research in this domain was based on machine learning techniques to classify movie reviews as positive or negative. According to the study, sentiment classifiers operate as a particular type of text classification task and use computational techniques to allow users to extract opinions from texts. This study further developed the idea by providing a comprehensive overview of sentiment analysis methods and specific applications in several disciplines. However, lexicon-based methods have been further improved with the advent of large-scale sentiment dictionaries. A well-known resource is National Research Council (NRC), which maps words to several emotional categories, including joy, anger, sadness, and fear [6]. This lexicon gained wide popularity in the context of sentiment analysis because it can successfully detect emotions expressed in text data.

The study found that sentiment analysis improved when feature engineering techniques were applied. Researchers have developed numeric conversion methods for text data using bag-of-words, TF-IDF, and n-gram models [4]. The system uses this information to identify both word frequencies and word relationships throughout the document. The development of these models depends on the algorithms. The system uses labeled training datasets to identify patterns, which are then used for prediction tasks. The new technology makes better predictions because it understands the data better than traditional systems, which depend on fixed regulations [7].

The effectiveness and performance of machine learning models depend on the expertise of their developers and the available training data. Machine learning models operate effectively when developers create high-quality feature sets and use appropriate data. Machine learning models require high-quality data, which should include notes for effective learning. Sentiment analysis has improved significantly because advancements in deep learning technology have gotten much better [8]. Deep learning technologies, such as Convolutional Neural Networks, Recurrent Neural Networks, and Long Short-Term Memory, are good at handling large amounts of text data. They get better at it.

They can figure out things about language independently. This is because of the ways of building deep learning models, such as the one called Bidirectional Encoder Representations from Transformers, or BERT for short.

Study Selection

There are various ways to accomplish this, such as observing how people communicate with each other and their behavior on social media sites. Our skills in this area have been enhanced, but it is not possible to identify whether a person is using sarcastic language or passing information that is culture-specific. Our review process began with reading several studies. This helped us identify key information. We read all available information. Our selection process involved choosing information that was necessary for our educational requirements. Computer language processing tools help us study media because they can understand what a person is saying. This system helps these tools monitor the information shared on social media sites. Online communication patterns help identify new trends in communication that were previously unknown. We surveyed tools that help users of instant messaging services understand the language being spoken by another user [6]. The research methodology consists of several stages, commencing with a literature search, followed by selection criteria, data extraction, and finally, a stage of comparative analysis. The methodology was designed to include the best literature in the study.

It aims to identify major trends, techniques, and research in the domain of social media sentiment analysis. The first step of the methodology was to conduct an extensive literature search, which required access to various academic databases. Prominent academic databases, such as ScienceDirect, SpringerLink, ACM Digital Library, and Google Scholar, were considered for retrieving research papers from various research publications. These databases were considered because they provide access to a wide variety of research papers published in journals and conferences related to artificial intelligence, machine learning, NLP, and data mining. Using multiple databases will cover all research contributions, thereby avoiding the chances of missing important literature [9]. To fetch relevant publications, various keywords and search queries were formulated based on the research topic. The main keyword was “sentiment analysis,” “opinion mining,” “natural language processing,” “social media analytics,” “machine learning for sentiment classification” and “deep learning in NLP.” [10]. The formulated keywords were then combined using the Boolean search operators “AND” and “OR” to narrow down the search results and make the search results more relevant to the research topic. For instance, the search queries were formulated as “sentiment analysis and social media.” “NLP and opinion mining.” “Deep learning and sentiment classification” to fetch research papers that directly deal with the research objectives of this review [11].

Literature Search Strategy

The literature search strategy for this systematic review is characterized by a meticulous process, where the goal is to comprehensively explore the developments of sentiment analysis, beginning with conventional lexicon-based techniques, through various methodological comparisons, benchmark studies, and even gaps such as sarcasm detection and multilingualism, to finally arrive at cutting-edge techniques such as BERT and GPT, which utilize transformers. Through various data sources, such as Twitter, Reddit, and e-commerce sites, this literature search strategy is geared towards obtaining a wide scope of how researchers address the special noise, brevity, and lexical characteristics of social media data. This systematic review is geared towards revealing emerging trends, such as multimodal fusion and explainable AI, while providing a transparent framework for assessing the status of sentiment analysis techniques [12].

Data Extraction and Analysis

A systematic method of data extraction and analysis was used by the researchers to obtain research studies for the purpose of carrying out research analysis. The research method was specifically designed to find all high-quality research articles that contained the latest research information on the topics of sentiment analysis, machine learning, and social media analytics. The keywords were searched by combining them with Boolean operators, which are specific words that are used to join keywords. These keywords included various terms such as “sentiment analysis,” “opinion mining,” “natural language processing,” “social media analysis,” “machine learning,” “deep learning,” “text classification,” “Twitter sentiment analysis,” “Facebook sentiment analysis,” and “social media data mining.” Boolean operators, such as AND, OR, and NOT, are also included in these keywords, which are essential in finding specific keywords that match their research. The keywords were as follows: “sentiment analysis and social media and natural language processing,” “opinion mining or sentiment classification and machine learning,” and “deep learning and social media sentiment analysis.”

Social Media Data Characteristics:

There are a lot of things being shared through social media. Social media is a source of information about what, specifically, is shared by people [7]. Such information could help us better comprehend the feelings and opinions of people regarding certain things. Therefore, social media is a means for researchers to seek and observe the emotions of others about something through sharing short messages.

Social media data have certain characteristics that make them different from other text data. It is important for researchers to have a good understanding of social media data to create better instruments for determining the intent of the author of certain text data [4]. Knowledge of the identification of social media data gives researchers the ability to create better instruments for determining the intent of the authors of social media data.

DATA SOURCES

When individuals search for data to run a sentiment analysis, it is probable that they would perform a sentiment analysis of social media sites, where most of their needs are likely to be met.

Therefore, there are many social media sites with good APIs where the information will be available for free in almost real-time. In addition, they tend to provide information regarding what someone thinks or feels about an experience after they have had it [2]. However, the most suitable of these two social media sites for performing a sentiment analysis is probably Twitter, based on the fact that it lends itself better to sentiment analysis than Facebook.

Twitter is a good example of a site used for sentiment analysis because it has an open API, and it is easy to find on Twitter because Twitter users have a limited amount of space with which to post their thoughts. So, most Twitter users use Twitter to be very direct and clear in their posts [4]. Twitter is home to many users who use hashtags, mentions, and emojis to convey a large amount of information in a relatively small number of characters or words.

Facebook, however. However, this is not the case for performing sentiment analysis. While it is also a site where the great majority of its users will express their thoughts or feelings about brands, products, and the news, users also generally tend to post a significantly larger amount of content on Facebook compared to Twitter; therefore, sentiment analysis is significantly more complicated (Figure 1).

Nature of Social Media Data

The content consists of short messages in informal language. However, it lacks proper organization. In addition, no one cares about grammar and spelling. Many slang and internet abbreviations are used. The use of the abbreviations “LOL,” “OMG,” and “By The Way (BTW)” is quite common. NLP is a problem in this type of writing because proper textual content is required.

Emojis and emoticons are different forms of visual communication. The visual component in the form of emojis and emoticons gives a cute look and displays the emotions expressed by a person [8]. A single emoji may display the emotions expressed by a person better than the text written by the person. The different meanings expressed through emojis are more than those of textual content.

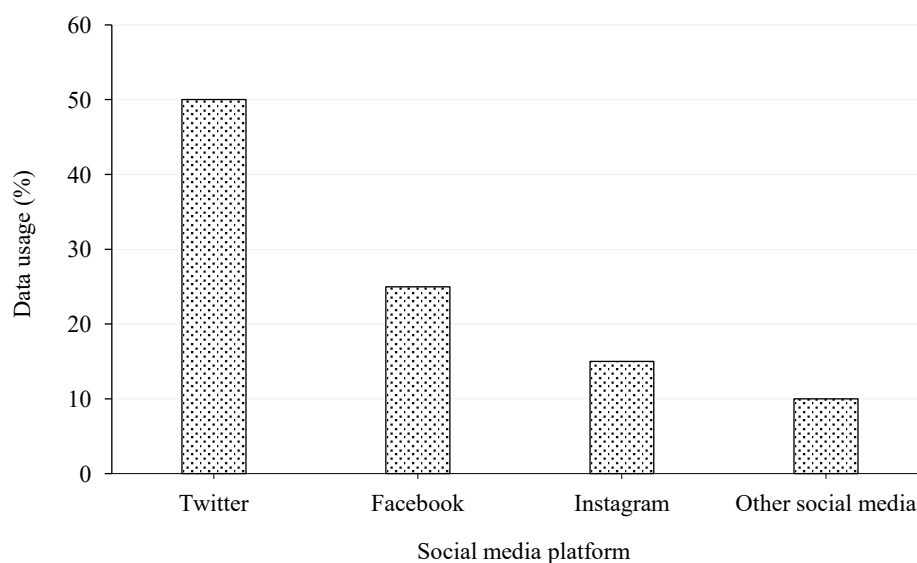


Figure 1. Distribution of social media datasets utilized in sentiment analysis, illustrating the percentage prevalence of each platform across surveyed literature.

Hashtags are used by many internet users. Hashtags help organize things and help others find things. The use of the hashtag “#fail” and “#Love it” helps a person express their emotional state regarding a particular topic. The use of mentions increases when a person wants to identify someone and wants others to join the conversation.

PREPROCESSING

The purpose of performing the sentiment analysis on the text data of social media is to be more important. It is required that the text data of social media be preprocessed using the appropriate text data preprocessing techniques. Considering the informal nature of the text data on social media. It is also important to highlight that the purpose of using the appropriate techniques for preprocessing the text data of social media is to ensure that the text data of social media are cleaned and normalized, maintaining the required information for sentiment analysis.

The first step of the text data preprocessing technique is the tokenization of text data, which can be achieved by splitting the text data into the required elements of the text data. This would allow the NLP Algorithm to split the text data into the required elements, thereby facilitating text data processing [7].

To normalize the text data, the text data must be processed using three techniques: converting the text data to lowercase, eliminating spelling mistakes, and eliminating word variations from the text data.

RECENT TRENDS AND ADVANCEMENT

The capabilities of sentiment analysis systems in terms of both social media data analysis and general application cases now include more advanced features through deep learning, language comprehension algorithms, and the ability to analyze multiple types of data simultaneously. Traditional lexical-based approaches to performing sentiment analysis using standard machine learning methods have now been supplemented by a variety of new techniques and methods for performing sentiment analysis of even very complex linguistic patterns and meanings from extremely large amounts of real-time user-generated content (i.e., social media). This section summarizes some of these new trends and innovations in sentiment analysis of research efforts [2].

CONCLUSION

Throughout this review, we concluded that the evolutionary journey of technology plays a vital role in the process of sentiment extraction of people’s opinions from different social media platforms and several human behaviors in day-to-day life. The early lexicon-based approach was used. The methods used for sentiment analysis have changed significantly. This change began a while ago and involved a list of words with feelings associated with each word. The first ones did not really learn; they were more about matching words with feelings. When the tone changed depending on the situation, they did not perform well. This is because they were based on rules, which were not flexible.

Something then came along, and people began to look at things in a different way, depending on what happened. This is when decision trees and vectors came in, and they were used in life situations where the feelings were known.

The importance of context is what makes a difference. Individuals do not necessarily express themselves in the way they intend, and they usually do not intend to do everything they say. When someone tweets something like “Oh fantastic, it’s Monday again,” he or she is not necessarily excited about it, and current sentiment engines are now finally getting this idea.

One major change that can be identified is that sentiment analysis has changed from being rigid to being more flexible. The change in sentiment analysis occurred in a specific manner. The original method of conducting sentiment analysis is no longer used. It is more about being able to change depending on the situation.

Future Scope

With the continuity of advancement in the respective field of sentiment analysis, several areas have been examined for future research to improve the precision of sentiment detection.

1. Handling linguistic complexity
2. Real-time scalability
3. Refinement of hybrid methodologies
4. Contextual sarcasm and irony detection
5. Advanced multimodal analysis
6. Linguistic adaptability
7. Visual sentiment integration

Acknowledgement

The authors thank the Department of Computer Science and Engineering at Greater Noida Institute of Technology, where the right mix of space, tools, and learning made this review possible. Support did not stop there - guidance from teachers and advisors played a quiet but steady role throughout the work. Their insights shaped ideas; questions grew sharper because of them; each comment left its mark, without fanfare. Even when things slowed down, this backing never faded.

We are grateful to every researcher whose insights into sentiment analysis helped shape this study. Their contributions, drawn from the studies carefully examined here, anchor the direction taken. Without these earlier efforts in NLP, this exploration would lack footing. Each citation reflects a piece of the foundation built by others.

Disclosure

1. No specific funding was received from public, commercial, or non-profit sources for this study.
2. The authors have no conflicts of interest.
3. This study is entirely based on previously published studies, with no new human or animal experiments.
4. All data were derived from the existing literature in the public domain without recourse to original data collection.

REFERENCES

1. Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. *Comput Intell.* 2006;22(2):110-25. doi:10.1111/j.1467-8640.2006.00277.x.
2. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. *Comput Intell.* 2013;29:436–465. doi:10.1111/j.1467-8640.2012.00460.x.
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1, Long and Short Papers; 2019.* p.4171–4186. doi:10.18653/v1/N19-1423.
4. Rosenthal S, Farra N, Nakov P. SemEval-2017 task 4: Sentiment analysis in Twitter. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); 2017.* p.502–518. doi:10.18653/v1/S17-2088.
5. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; 2013.* p.1631–1642. doi:10.18653/v1/D13-1170.
6. Liu B. *Sentiment analysis.* Cambridge: Cambridge University Press; 2020. Available from: <https://www.cambridge.org/core/books/abs/sentiment-analysis/sentiment-analysis/0AB94D9F40A0402AB6EA566CB5776D30>.
7. Schouten K, Frasincar F. Survey on aspect-level sentiment analysis. *IEEE Trans Knowl Data Eng.* 2016;28:813–830. doi:10.1109/TKDE.2015.2485209.

-
8. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv [Preprint]. 2013. Available from: <https://arxiv.org/abs/1301.3781>
 9. Bari P, Bedi G, Joshi K, Jawale A. Why transformers outperform LSTMs: A comparative study on sarcasm detection. *J Artif Intell.* 2025;7:499–508. doi:10.32604/jai.2025.072531.
 10. Yoshikawa T. *Foundations of Robotics: Analysis and Control*. Cambridge (MA): MIT Press; 1990. doi:10.7551/mitpress/3074.001.0001.
 11. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. arXiv [Preprint]. 2019. Available from: <https://arxiv.org/abs/1906.08237>
 12. Dediu AH, Matos M, Martín-Vide C. Natural language processing, moving from rules to data. In: *International Conference on Theory and Applications of Models of Computation*. Cham: Springer International Publishing; 2017. p.24–38.