E-ISSN 2250-0944
ISSN 2250-1150
doi: 10.31838/ijprt/14.01.08

Research Article

Navigating Sentiment Analysis Horizons: A Comprehensive Survey on Machine Learning Approaches for Unstructured Data in Medical Sciences and Science and Technology

P.SUGANYA[1], G.VIJAIPRABHU[2], G. SIVAKUMAR[3], K. SATHISHKUMAR[4]
[1]Ph.D Research Scholar, PG and Research Department of Computer Science, Erode Arts and Science College(Autonomous) Erode, Tamilnadu,India.
[2]Assistant Professor, PG and Research Department of Computer Science, Erode Arts and Science College (Autonomous) Erode, Tamilnadu,India.
[3]Assistant Professor, Department of CS – PG, K.S.Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamilnadu,India.
[4]Assistant Professor, PG and Research Department of Computer Science, Erode Arts and Science College (Autonomous) Erode, Tamilnadu,India.
Received: 18.09.23, Revised: 26.10.23, Accepted: 09.11.23

ABSTRACT
The increasing field of sentiment analysis on unstructured data has become a focal point of research, witnessing a proliferation of machine learning techniques. This extensive survey investigates into the methodologies embraced by researchers across diverse domains, spotlighting the pivotal role played by automatic feature learning and word embedding models in the booming success of sentiment analysis approaches. The exploration of sentiment analysis techniques, the survey unravels the workings of Support Vector Machine (SVM), Naive Bayes (NB), Artificial Neural Networks (ANN), Decision Trees (DT), K Nearest Neighbour (k-NN), Random Forest (RF), and metaheuristic optimization algorithms, elucidating their time complexities, advantages, and limitations. By synthesizing the challenges faced by researchers, the survey not only offers prominent insights but also depicts the course for future investigations, presenting an open issue in the sentiment analysis. The discourse extends beyond theoretical considerations to practical applications, evaluating the performance of sentiment analysis techniques across a spectrum of real-world datasets. As a comprehensive resource, this survey provides researchers and practitioners with a understanding of the evolving paradigm, fostering informed decision-making and inspiring future innovations in sentiment analysis on unstructured data within the paradigm of machine learning.

Keywords: Sentiment analysis, machine learning, unstructured data, lexicon, hybrid model, and linguistic structure, Medical Sciences.

INTRODUCTION

Sentimental Analysis (Liang-Chu Chen et al. 2019) is the task of determining whether the given piece of text is positive, negative, or neutral. A sentimental analysis system integrates natural language processing and machine learning techniques to assign a probabilistic score to each object, topic, sentiment, theme, or class. Data analytics industries often incorporate third-party application interfaces into their customer experience management system to offer users insights to their end-users. The sentimental analysis conducted in the documents is a straightforward process. Initially, each text document is broken down into different components (sentences, phrases, tokens, and parts of speech). The next step is to identify the sentiments present in each component. After that, the sentiment score is assigned to each component, and these scores are integrated with conducting a multi-layered sentimental analysis.

Sentimental analysis is mainly based on polarity (positive, negative, and neutral), emotions (happy, sad, angry, depressed, etc.), necessity (needed/ not needed), and intentions (interested/not-interested). Based on the customer feedback and queries, one can design a sentimental analysis system to match their needs. Both sentimental analysis and opinion mining are similar in many ways. The sentimental analysis identifies the emotions present in the natural language text, whereas the opinion mining (Liu Bing et al. 2012) extracts sentiments from the text. The textual information can be classified into two types one based on facts and the other based on

opinionated information. Both the objective and subjective sentences fall under facts that contain clear opinions, incidents, and views about different products.

The increasing amount of Web 2.0 devices resulted in a massive amount of data generated daily. Hence sentimental analysis serves as an important tool to derive insights from user-generated data. Different approaches utilizing the keyword, lexicon, and machine learning techniques have been proposed for sentimental analysis, and they achieved the standard performance. This chapter analyzes the current performance of different techniques and the various challenges that need to be addressed to present an efficient sentimental analysis tool. The results provided by different techniques for various datasets are also explored. The recent multimodal sentimental analysis papers were also reviewed. The different application areas (Medical field, Financial Sector, Sports, Government, Tourism, Politics, Marketing, and sales) where the sentimental analysis techniques offer major benefits are also explored. The open issues that serve as guidance for aspiring researchers are also discussed at the end of this article.

Sentiment Analysis Technique

The sentimental analysis process is mainly initiated by collecting data from different sources such as social media to identify the different polarities present in the text. If the sentimental analysis approach aims to identify the related keywords, a keyword search is initiated. The sentimental analysis system is mainly constructed using different techniques such as keywords, lexicon, machine learning, and deep learning techniques (Shahid Shayaa et al. 2018). The taxonomy of the sentimental analysis approach is presented in Figure 1, and the application of these techniques is presented in these subsections.

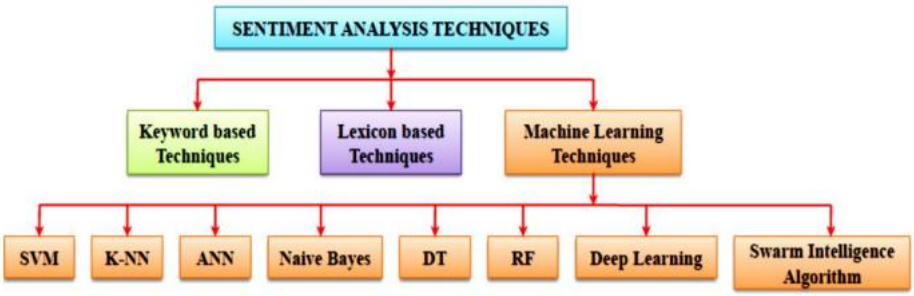

Figure 1: Taxonomy of Sentimental Analysis Techniques

When conducting sentimental analysis on a textual dataset, the words that accurately describe the entire document are found, which is useful when reducing dimensionality, computational time, removing unwanted/ redundant terms, and improving accuracy and performance. The significant terms extracted from the entire document include a description of the content; these words/terms are referred to as keywords. The method used to extract them is referred to as the keyword extraction technique. Monali Bordoloi et al. (2020) presented a co-occurrence graph-based statistical approach to find the global rank of keywords. A new weighting technique is used to improve the standard Node and Edge ranking technique. A keyword can also exhibit bipolarity based on the problem at hand. The author proposed a novel graph-based algorithm that gives a higher priority to the important keyword when compared to the least significant one because the keywords play a prominent role in determining the polarity of the text. Baumgarten et al. (2013) presented a keyword-based sentimental mining approach to analyze the tweets present on Twitter. The authors mainly use the keyword-based classifier to perform sentiment mining in short messages, and the approach can be automatically extended for messages with multiple dimensions. The main challenge encountered here is the non-trivial problem of extracting specific features/aspects from the short messages.

The main aim of the lexicon-based techniques is to identify the sentiments present in documents or sentences by analyzing the sentimental aspects present in the user's writing. The sentiment words are used to express the sentiments such as "happy" (positive)," wow" (positive), "shit" (negative), and "terrible" (negative). These words are known as opinion lexicon or SentiWordNet. Since the lexicon-based approaches use supervised learning, they require an external knowledge source for training. It can

be in the form of a labeled lexicon that contains a polarity associated with each sentimental word. The lexicons are also implemented in developing labeled training data for the machine learning classifier. The polarity is mainly expressed using a numerical value that indicates how strong a particular word is associated with a positive and negative polarity.

Even though lexical based techniques are efficient in conducting the sentimental analysis, it often ignores the contextual information associated with the sentence. To overcome this problem, Minghui Huang et al. (2020) developed a lexicon-based attention mechanism for their Sentiment Convolutional Neural Network (SCNN) to analyze both the sentiments and contextual information derived from the sentiment words. The contextual information is mainly captured from the word embeddings, and it is a prominent indicator of sentiments to make effective predictions. This technique offers accuracy, precision, recall, and F1-Measure 88.4%, 88.9%, 88.9%, and 88.9%.

Kristína Machovaet al. (2020) applied the lexicon approach for automatic labeling to overcome the complexities associated with ambiguous and subjective manual labeling. To optimize the lexicon labeling approach, Particle Swarm Optimization (PSO) technique is used. The PSO optimizer repeatedly labels every word present in the lexicon and evaluates the opinion classification approach after every optimal label for the words present in the lexicon is identified. This hybrid approach can classify more than 99% of text accurately and achieve better results than the traditional lexicon-based approaches. Mahmoud Al-Ayyoub et al. (2015) presented a lexicon-based sentimental analysis approach for Arabic tweets. The main complexity of the polarity classification is the Arabic language's intricate structure and fewer datasets present for processing. The authors mainly used the lexicon approach to build a large sentimental lexicon and a sentimental analysis engine.

Machine learning algorithms serve as useful in handling massive datasets and solving real-world problems. They are classified into two groups, namely supervised and unsupervised learning algorithms. The supervised learning algorithm mainly inputs the labeled dataset, and the accuracy of the outcomes is evaluated using the training dataset. As a consequence, supervised learning is best suited to problems that have a variety of available reference points or ground truth to train the algorithm with Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), Artificial Neural Network, Naïve Bayes, Decision tree, and random forest are some examples of supervised learning. An unsupervised model takes the unlabelled data as an input and extracts the features and patterns present in it without using any external support(manual intervention) and on its own. It resembles a black box structure of processing. A deep learning algorithm is given a dataset with no specific instructions about what to do with it in unsupervised learning. The training dataset is a collection of examples with no particular desired outcome or right answer. The neural network then attempts to automatically find meaning in the data by extracting useful features and analyzing the data's structure. Deep learning techniques mainly use unsupervised learning.

Amita Jain et al. (2020) designed a hybrid Senti-NSet PSO approach by integrating the PSO algorithm with the Neutrosophic set technique. When the input file size exceeds 25KB, this method is appropriate. The main aim of the SentiNSet PSO is to handle the complexities associated with large texts and to identify the sentiment polarity with higher accuracy. The neutrosophic method is centered on a three-valued tuple that includes truth, fallacy, and indeterminacy. The neutrosophic technique is used to optimize the PSO to derive the overall sentiments. Sayar Singh Shekhawat et al. (2020) developed a hybrid metaheuristic-based clustering technique by integrating the spider monkey optimization with the k-means clustering algorithm. The hybrid Spider Monkey Optimization (SMO) with kMeans clustering(SMOKM) algorithm is mainly introduced to achieve the optimal cluster heads for the dataset. Initially, the k-means algorithm is used to generate k clusters, and the SMO algorithm is used to optimize the cluster head selection process. The SMOKM technique offers an accuracy of 99.98% on the Twitter dataset and an accuracy of 88.93% on the Twitter sanders dataset. Abdulaziz Alarifi et al. (2020) developed a big data-based sentimental analysis approach by integrating the greedy feature selection with a Cat Swarm Optimization (CSO) based LSTM Neural Network (CSO-LSTMNN). The crucial features are selected using a greedy algorithm which is then processed using the CSO-LSTMNN approach. The main aim of this algorithm is to overcome the problems such as outdated memory and diversity loss. The CSO algorithm is applied mainly to enhance the error rate, accuracy, precision, and recall of the LSTMNN.

Comparative Analysis on Sentiment Analysis Technique

The Support Vector Machine (SVM) excels with structured data and high generalization but suffers from increased training time with larger datasets. Fine-tuning parameters like C and gamma can be challenging. Naive Bayes (NB) is useful for multi-class prediction but assumes attribute independence. Artificial Neural Networks (ANN) handle incomplete knowledge but require parallel processing and are prone to overfitting. Decision Trees efficiently handle non-linear datasets but are sensitive to slight data changes. K-NN is easy to implement but struggles with high-dimensional datasets. Random Forest reduces overfitting but demands computational power. Metaheuristic Optimization Algorithms are versatile but time-consuming. The prominent aspects are given in Table 1.

Table 1. Comparative Analysis on Sentiment Analysis Technique

Table 1: Time Complexity, Advantages, and Disadvantages of the Machine Learning Algorithms

Algorithm Name	Advantages	Disadvantages	Time Complexity	Parameter Description
Support Vector Machine	Works well with structured and semi-structured data and has high generalization performance	As the size of the dataset increases, the training time also increases, and fine-tuning the parameters such as C and gamma is hard	O(s 2)	s-total number of samples present in the dataset
NB	Widely used for multi-class prediction problems and utilizes much less data	It assumes that all attributes present are mutually independent, and this scenario is not possible in realtime	O(sf)	f-total number of features s-total number of samples
ANN	Ability to handle incomplete knowledge	It requires processors with parallel processing power, and the model is vulnerable to overfitting	O(efsn)	e-Total number of epochs f-total number of features s-total number of samples present in the dataset n-total number of neurons
Decision Tree	It handles non-linear datasets efficiently	A slight change in the data results in a huge variation in the decision tree structure	O(f s2)	f-total number of features s-total number of samples
K-NN	It does not involve training for predictions and is easier to implement	t does not scale well for highdimensional and large datasets and is sensitive to noise, missing values, and outliers.	O(s)	s-total number of training samples
Random Forest	Reduces the overfitting problem in decision trees	Needs higher computational power and training time	O(Afslogs)	s-total number of samples f-total number of features A-total number of trees

	and improves their accuracy			
Metaheuristic Optimization Algorithm	Easier to implement and can handle different objectives and constraints	Consumes a higher amount of time, and the delay can affect the system negatively	O(fsips)	s-total number of samples f-total number of features i-Total number of iterations ps- the size of the population

Sentiment analysis techniques have witnessed significant advancements in extracting sentiments from textual data, particularly through the exploration of various methodologies such as lexicon-based approaches, machine learning algorithms, and hybrid models. However, a notable research gap lies in the limited consideration of contextual information within lexicon-based techniques. While these approaches efficiently identify sentiments, they often fall short in capturing the nuanced contextual nuances associated with sentences. For instance, the work of Minghui Huang et al. (2020) introduces a Sentiment Convolutional Neural Network (SCNN) with a lexicon-based attention mechanism, addressing this gap by enhancing lexicon techniques to incorporate and leverage contextual information for more nuanced sentiment analysis. Further research is warranted to explore and develop advanced lexicon-based models that effectively integrate contextual insights, thereby improving the accuracy and comprehensiveness of sentiment analysis, especially in scenarios where contextual nuances play a crucial role.

Another significant research gap pertains to the challenges posed by sentiment analysis in languages with intricate structures, exemplified by the Arabic language. Although Mahmoud Al-Ayyoub et al. (2015) presented a lexicon-based sentiment analysis approach for Arabic tweets, the intricacies of the language's structure and the scarcity of datasets present ongoing challenges. There is a need for more in-depth exploration and development of techniques specifically tailored to overcome language-specific complexities, ensuring accurate sentiment analysis in languages like Arabic. Bridging this research gap would involve a comprehensive understanding of linguistic nuances and the creation of robust sentiment analysis models that are adaptable to diverse linguistic structures.

In the domain of machine learning algorithms for sentiment analysis, a notable research gap exists in optimizing hybrid approaches to handle large datasets effectively. While various hybrid models, such as the Senti-NSet PSO approach (Amita Jain et al., 2020), have been proposed, there remains room for further exploration to address the challenges associated with increasing file sizes. As datasets grow in size and complexity, the efficiency and accuracy of sentiment analysis models become paramount. Researchers should focus on refining and optimizing hybrid machine learning approaches to accommodate the demands of extensive textual datasets. This involves investigating strategies to enhance the scalability, processing speed, and overall performance of sentiment analysis models, ensuring their applicability and effectiveness in real-world scenarios with substantial data volumes.

Addressing these research gaps is crucial for the continued advancement of sentiment analysis techniques. Future research endeavors should aim to enhance lexicon-based approaches by incorporating contextual information, delve deeper into the complexities of sentiment analysis in languages with intricate structures, and optimize hybrid machine learning models to effectively handle the challenges posed by large and complex datasets. By addressing these gaps, researchers can contribute to the development of more accurate, robust, and adaptable sentiment analysis methodologies with broader applicability across diverse linguistic and data-intensive contexts.

Application on Sentiment Analysis Technique

Nalini Chintalapudi et al. (2021) presented a novel text mining application for sentimental analysis using seafarer health records. Seafarers are highly prone to accidents and health issues. The researchers utilized the vast textual data generated from digital health systems such as Electronic Health Records(EHR), medical prescriptions, and observation notes. This information collection leads to improved quality of treatment, low medical errors, and minimal costs. Patient statements about the explanation of illness are also positively matched with the diagnosis by a doctor, using the results obtained by the lexicon sentimental analysis and NB technique. Sasikalaet al.(2021)presented a Deep

Neural Network(DNN) technique to perform sentimental information on the healthcare information obtained from the social media text.

Young-Eun Park et al. (2020) presented a novel predictive data mining technique to predict the growth of the financial sector of Saudi Arabia using social media textual data. The main aim of this predictive framework is to evaluate client choices systemically and feedback for financial companies and identify key demand factors. Douglas H Silva et al.(2020) analyzed the sentiments present in the soccer game messages obtained from social media users. People are always interested in the insights gained about a particular player or user when it comes to sports data. The authors used a phrase-level metric known as word dictionary to retrieve features related to achieving this objective. They were also considerate about the gender-related features.

Rodrigo Sandoval-Almazan et al. (2018) analyzed the user sentiments regarding the local government campaign held in Mexico. The results illustrated that the political party with the most negative comments won the election, while the political party with the most positive comments failed. Alireza Alaei et al. (2019) used a big data approach to analyze the semantic relationship and meanings in tourist reviews by performing sentimental analysis on social media content. They also looked at how the big data paradigm can be applied to tourism to understand the nature of massive distributed datasets better.

Yan Liu et al. (2020) investigated the end user's sentimental values for brand attachment using three features: symbolic, functional, and aesthetic values. Pushpendu Rakshit et al. (2021) conducted a sentimental analysis on Indian big billion days of Flipkart and amazon using a people-based personalized approach to investigate the probability of the marketer who wins the millennial consumer sentiment. Using the analytical-based statistical tools on the millennial samples (5000 tweets) from the Twitter platform regarding the Amazon and Flipkart big billion day sales 2019, the sentimental analytics was conducted. To analyze the subject of sentiment, subjectivity, and polarity, a rule-based approach is also proposed. To identify the different polarity classes, they used fine-grained analysis and aspect-based sentimental analysis. Compared to Amazon, Flipkart received a higher number of negative comments because its services and product quality were deemed inferior by customers. Karthik et al. (2021) presented a product recommendation system that recommends products to the users based on their sentimental score of the product obtained using the fuzzy logic technique. The fuzzy rules are integrated with ontology techniques to yield efficient decision-making and increased accuracy.

CONCLUSION

Sentimental analysis using machine learning techniques has attracted a large number of researchers towards this area which results in a massive amount of sentimental analysis models being proposed. The automatic feature learning techniques and the word embedding models used by various authors have resulted in the great success of these approaches. This chapter provides the different techniques used by various authors to conduct sentimental analysis in different fields. It also discusses the advantages and limitations each researcher faced while developing the sentimental analysis model. The time complexity of different techniques such as Support Vector Machine, NB, ANN, Decision Tree, k-NN, Random Forest, and metaheuristic optimization algorithm, along with their advantages and disadvantages, are also provided. The different challenges faced by the existing literary works that need to be analyzed in the future are presented in the open issues section. The various applications of sentimental analysis techniques and their performance in various real-world datasets are also investigated.

REFERENCE

1. Alaei, Ali Reza, Susanne Becken&BelaStantic 2019, ‘Sentiment analysis in tourism: capitalizing on big data’,Journal of Travel Research, vol. 58, no. 2, pp. 175-191.
2. Alarifi, Abdulaziz, AmrTolba, Zafer Al-Makhadmeh&Wael Said, ‘A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks’, The Journal of Supercomputing, vol. 76, no. 6, pp. 4414-4429.
3. Baumgarten, Matthias, Maurice D Mulvenna, Niall Rooney & John Reid 2013, ‘Keyword-based sentiment mining using twitter’, International Journal of Ambient Computing and Intelligence (IJACI), vol. 5, no. 2, pp. 56-69.
4. Bordoloi, Monali&Saroj Kr Biswas 2020, ‘Graph based sentiment analysis using keyword rank based polarity assignment’, Multimedia Tools and Applications, vol. 79, no. 47, pp. 36033-36062.

5. Chen, Liang-Chu, Chia-Meng Lee & Mu-Yen Chen 2019, 'Exploration of social media for sentiment analysis using deep learning', Soft Computing, pp. 1-11.
6. ChintalapudiNalini, GopiBattineni, Marzio Di Canio, Getu Gamo Sagaro & Francesco Amenta 2021, 'Text mining with sentiment analysis on seafarers' medical documents', International Journal of Information Management Data Insights, vol. 1, no. 1, pp. 100005.
7. Huang, Minghui, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard KM Poon, & Fu Lee Wang 2020, 'Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis', IEEE Transactions on Affective Computing.
8. Huang, Minghui, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard KM Poon, & Fu Lee Wang 2020, 'Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis', IEEE Transactions on Affective Computing.
9. Jain, Amita, Basanti Pal Nandi, Charu Gupta & Devendra Kumar Tayal 2020, 'Senti-NSetPSO: large-sized document-level sentiment analysis using Neutrosophic Set and particle swarm optimization', Soft Computing, vol. 24, no. 1, pp. 3-15.
10. Karthik, RV & Sannasi Ganapathy 2021, 'A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce', Applied Soft Computing, vol. 108, p. 107396.
11. Liu, Bing 2012, 'Sentiment analysis and opinion mining', Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167.
12. Liu, Yan, Yan Kou, Zhenzhong Guan, JiaJing Hu & Bo Pu 2020, 'Exploring hotel brand attachment: The mediating role of sentimental value', Journal of Retailing and Consumer Services, vol. 55, pp. 102143.
13. Mahmoud, SafaBaniEssa&IzzatAlsmadi 2015, 'Lexiconbased sentiment analysis of arabic tweets', International Journal of Social Network Mining, vol. 2, no. 2, pp. 101-114.
14. Moraes, Rodrigo, João Francisco Valiati & Wilson P Gaviã O Neto 2013, 'Document-level sentiment classification: An empirical comparison between SVM and ANN', Expert Systems with Applications, vol. 40, no. 2, pp. 621-633.
15. PARK, Young-Eun & Yasir Javed 2020, 'Insights Discovery through Hidden Sentiment in Big Data: Evidence from Saudi Arabia's Financial Sector', The Journal of Asian Finance, Economics, and Business, vol. 7, no. 6, pp. 457-464.
16. Rakshit, Pushpendu, Pramod Kumar Srivastava, MohdAfjal & Shailendra Kumar Srivastava 2021, 'Sentimental Analytics on Indian Big Billion Day of Flip Kart and Amazon', SN Computer Science, vol. 2, no. 3, pp. 1-8.
17. Sasikala, D 2021, 'Effective Deep Neural Network Method based Sentimental Analysis for Social Media Health Care Information', Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 10, pp. 1883-1889.
18. Shayaa, Shahid, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai, YeongWai Chung, Arsalan Zahid Piprani & Mohammed Ali Al-Garadi 2018, 'Sentiment analysis of big data: Methods, applications, and open challenges', IEEE Access, vol. 6, pp. 37807-37827.
19. Shekhawat, Sayar Singh, Sakshi Shringi & Harish Sharma 2020, 'Twitter sentiment analysis using hybrid Spider Monkey optimization method', Evolutionary Intelligence, pp. 1-10.
20. Silva, Douglas Henrique, Renata Lopes Rosa & Demostenes Z Rodriguez. 2020, 'Sentimental Analysis of Soccer Games Messages from Social Networks using User's Profiles', INFOCOMP Journal of Computer Science, vol. 19, no. 1.