

Review on Sentiment Analysis of Health Related Dataset based on Machine Learning Techniques

G. Kavitha¹ and Dr. M. Vanitha²

¹Ph.D Research Scholar, Alagappa University, Karaikudi
Email: gkaviajan@gmail.com

²Assistant Professor, Alagappa University, Karaikudi
Email: vanitham@alagappauniversity.ac.in

Abstract— The interpretation and categorization of sentiments and opinions from the text is known as sentiment analysis. Due to the rapid propagation of minor health-related events, it is very difficult to extract information from unorganized narratives. Despite neural network techniques' effectiveness in enhancing learning structural dependability, when using less valuable feature sets, they provide insufficient accuracy in the categorization of bio-medical sentiments. They have no discriminating potential as a result. In this study, we analyze the research areas of sentiment analysis which implements machine learning methods on the dataset from large scale patients' narratives, social media and biomedical dataset.

Index Terms— Natural Language Processing, Sentiment analysis, Opinion Mining, Machine Learning and Text Mining

I. INTRODUCTION

Patient stories distributed on social media and health-related websites incorporate comments on a patient's wellbeing state as well as a riches of genuine information, like subtle elements of tests and medicines. Exploiting such data almost clients is key to the operation of numerous applications (such as recommender frameworks), within the overview investigations conducted by organizations, for occasion arranging of political campaigns [17].

Sentiment analysis (SA) is frequently utilized to get patient opinions approximately a certain thing or condition from web discussions, both positive and negative. Client estimation investigate empowers companies to distinguish how the common open feels approximately their products, administrations, and brands by means of online talks and surveys. Content categorization and normal dialect handling are basic for assumption investigation. That infers that, based on their reactions, ready to figure or categorize consumers' opinions.

This research addresses comprehensive comparison of estimation examination strategies within the literature, and an experimental study to assess the execution of machine learning models and related methods on datasets almost diverse subjects. The rest of this paper is organized as follows. Section 2 gives foundation information on this research area. Section 3 discusses related work. Section 4 depicts the commitment, taken after by the conclusion in Section 5.

II. BACKGROUND

A. Natural Language Processing (NLP)

A computer program's capacity to comprehend characteristic dialect, or human dialect because it is spoken and composed, is known as normal language processing (NLP). A portion of Artificial Intelligence (AI) incorporates it. It has various viable employments in a extend of businesses, counting corporate insights, look motors, and therapeutic investigate. NLP mixes factual, machine learning, and profound learning models with computational linguistics—rule-based displaying of human language [13]. With the utilize of these advances, computers are presently able to process human dialect within the frame of content or sound information and completely understand what is being said or composed, counting the speaker's or writer's intentions and estimation. It is greatly challenging to form program that dependably discovers the expecting meaning of content or voice information since human dialect is overflowing with uncertainty. Estimation Investigation is one of the key employments for NLP. The following Figure 1 shows the various NLP applications.

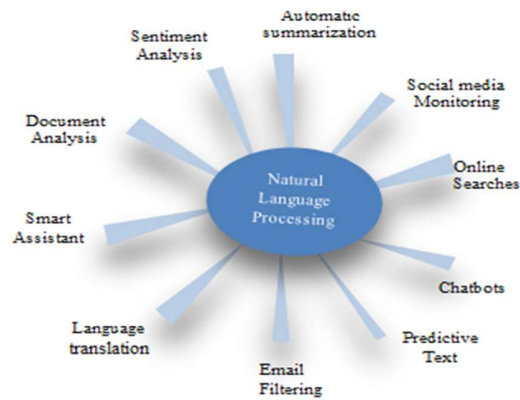


Figure 1. NLP Applications

B. Sentiment Analysis (SA)

Sentiment analysis is a technique in which a statement is broken down into smaller bits in order to discover how those parts interact with one another [4]. Typically, a piece of text serves as the input and the result of a sentiment classification model is the likelihood that the sentiment communicated is good, negative, or neutral. Sentimental analysis keep a variety of useful field areas in mind, such as item consumption price, competitive aptitude, calculation for product promotion, selection prediction [18, 24], investigation of resident associations, examination of banking system possibilities, and depiction of widespread diversions beginning in the industrial area.

The goal of sentiment analysis is to find out what individuals think about a particular topic, then to identify the described sentiments and, last, to classify them according to their dimensions, as illustrated in Figure 2. Machine learning methods participates a momentous character in SA.

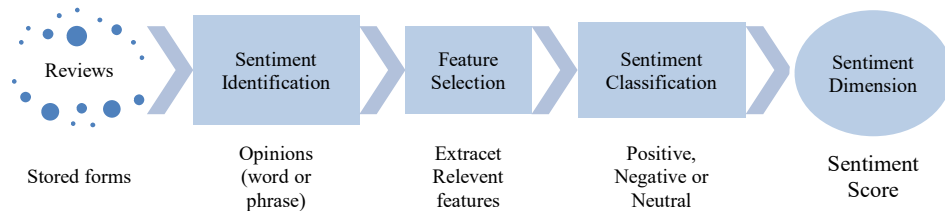


Figure. 2 Process of Sentiment Analysis

C. Machine Learning

A machine learning technology called sentiment analysis examines texts for polarity, ranging from positive to negative. Machine learning technologies can automatically learn how to recognize sentiment without human input by being trained with samples of emotions in text. Machines are programmed and trained using a variety of

methods and sophisticated algorithms to perform sentiment analysis. To ascertain the dimensions of textual data, a variety of classifiers, including Naive Bayes, maximum entropy, and support vector machines, are used [23] as shown in Figure 3. However, drafting explanatory notes collection is required, and this is especially true for difficulties involving many languages and different areas. A classifier designed for one domain may not perform better in another, just as in situations when many languages are present.

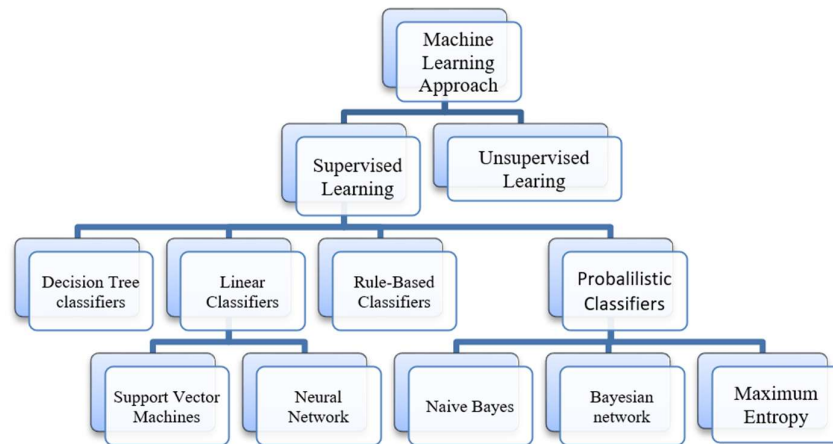


Figure.3 Types of Machine Learning Approach

D. Mining

The process of analysing patterns and learning from data is known as "data mining." It is a technique for extracting the important information from a data source and organising it logically. The different mining concepts are described in Figure. 4. Data mining is the process of obtaining knowledge from vast amounts of data using mathematical techniques and algorithms [19].

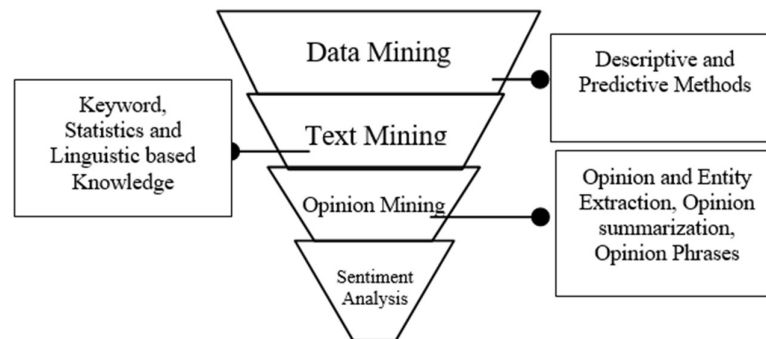


Figure. 4 Mining Concepts

Opinion Mining (OM) is the analysis of human behavior based on their textual contents. It is known as sentiment analysis as well. Sentiment analysis is broken down into three levels: aspect-based, sentence-based, and document-based. To complete these stages using the various OM approaches, including the Part-of-Speech (POS) extraction, opinion extraction, opinion categorization, and opinion summarization.

Text mining is the process of finding knowledge sources that comprise text or unstructured data. As a result, the primary distinction between text mining and data mining is that text mining uses unstructured data. The use of opinion mining in content analysis, mathematic operations, semantics, and the biometrics field is particularly beneficial. Scientifically, it is proven to dig up, classify, and count information in order to learn biased information as well as sentimental remarks. It is actually a component of NLP because it possesses a system's ability to recognize people's communication in the same way that it recognizes their written contents. Deriving the abstract subjects from writing content and looking for semantically based designs in textual form for this is a crucial issue in machine learning and natural language processing.

III. RELATED WORKS

This section describes a brief overview of sentiment analysis that had been discussed by different researchers work. Figure.5 shows the general overview related to sentiment analysis approaches according to their respective domains.

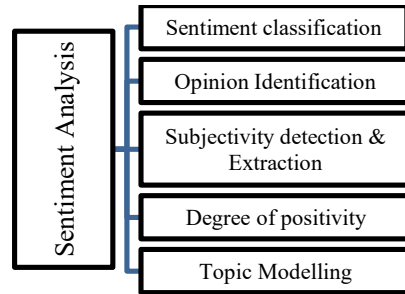


Figure. 5 Research Area of Sentiment Analysis

E. Sentiment classification

Sentiment classification is an automated approach that uses the thoughts expressed in a text to identify and categories emotions as positive, negative, or neutral. With the aid of Natural Language Processing (NLP), the type and degree of emotions expressed by the clients regarding company, goods, and services could ascertained. The reviews, comments, survey results, and other public opinions more quickly would be evaluated by using automated classification techniques. The following table 1. Summarizes the studies with different approaches in Sentiment Classification.

F. Opinion Identification

It aims to identify and classify particular viewpoints that are communicated through speech, text, or other media. It entails recognizing a broad spectrum of feelings, including more complex ones like enthusiasm or dissatisfaction, as well as more universal ones like happiness, sadness, wrath, fear, and surprise. In order to effectively infer opinions, they are detected through the use of deep learning, machine learning models, and multimodal analysis (which combine text, voice, facial expressions, etc.).

This process is done by lexical and machine learning approaches. Figure.6 depicts the steps of Lexicon based approach. A dictionary is built to hold the polarity values of lexicons for the lexical approach. When determining a text's polarity, the "overall polarity score" is obtained by adding the polarity scores of all the text's words, if they are listed in a dictionary. For instance, the text's overall polarity score rises when a lexicon matches a term that the dictionary indicates as positive. A text is categorized as positive if its total polarity score is positive; if not, it is categorized as negative [6].

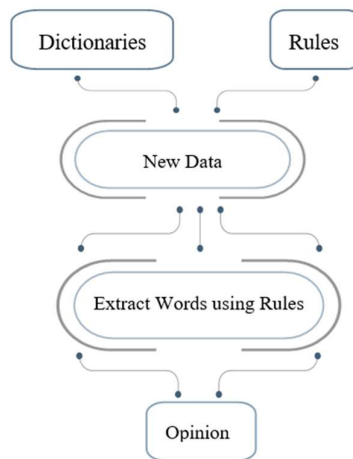


Figure 6. Lexicon based Approach

TABLE I. OVERVIEW OF STUDIES WITH DIFFERENT APPROACHES IN SENTIMENT CLASSIFICATION

Author	Year	Techniques	Contribution	Conclusion
Miftahul et. al., [16]	2022	TextBlob TF-IDF LinearSVC	Three distinct vectorization techniques (CountVec, TF-IDF, and Doc2Vec), three distinct vocabulary normalization procedures (potter stemming, lemmatization, and potter stemming plus lemmatization), and three distinct sentiment analysis techniques (VADER, TextBlob, and Azure). With a precision score of 0.96921 and the best model accuracy of 0.96752, the proposed experiment was successful.	The greatest model accuracy of 0.96752 was attained by LinearSVC employing TextBlob sentiment computation with TFIDF vectorization; this is greater than that of the Gradient Boosting Machine model classifier from the earlier research study. However, the accuracy of the classification model decreases when two vectorizations (CountVectorizer and LinearSVC) are combined; this may be because of the added complexity.
M. Arbane et al.[15]	2023	Bidirectional Long Short-Term Memory (Bi-LSTM)	Enhanced natural language processing approach for categorizing social media sentiment on COVID-19-related topics. Based on a Bi-LSTM approach, the suggested model is evaluated under four distinct conditions using actual datasets taken from the social media sites Reddit and Twitter.	Healthcare officials can use the suggested approach to better understand the needs and concerns of those impacted by the COVID-19 pandemic. Furthermore, official entities might effectively use the suggested methodology to lessen the impact of critical remarks.
D. C. Edara et al.[7]	2019	Long short-term memory (LSTM)	A distributed framework to analyse mood of cancer affected patients from various online cancer supporting communities. The corpus was constructed by collecting patient reviews from various domains using Twitter API.	The effectiveness of proposed framework was evaluated on the proposed dataset (corpus-1) and other two benchmark datasets like Health news Tweets (corpus-2) and Medical abstracts (corpus-3).
Arwa Alqarni and Atta Rahman[5]	2023	Convolutional neural networks (CNN) and bi-directional long short memory (BiLSTM) deep learning algorithms	Identifying and comparing the Arabic tweets sentiment analysis before and during the COVID-19 pandemic in Saudi Arabia.	The deep learning algorithms, CNN and BiLSTM, were used for analyzing Arabic sentiment. Moreover, the outcomes showed that the CNN algorithm is superior to BiLSTM, with 93% accuracy. Hence, the CNN model outperformed the BiLSTM by approximately 1%. So, a comparison has been made between proposed work and previous works in the same field of Arabic sentiment analysis related to COVID-19.

The following Figure. 7 show the method of Machine Learning approach. To construct a model for the machine learning approach, a set of labelled corpora is combined with a selection of feature vectors. After that, an untagged text corpus can be classified using the model. The choice of features in a machine learning approach determines how successful the categorization will be. Typically, feature vectors are selected from a range of unigrams (single words from a document) and n-grams (two or more words from a document in sequential sequence). The most widely used machine learning classification approaches include the Maximum Entropy algorithm, Naive Bayes, and Support Vector Machines (SVMs). The features used have a significant impact on how accurate the outcomes are for these algorithms.

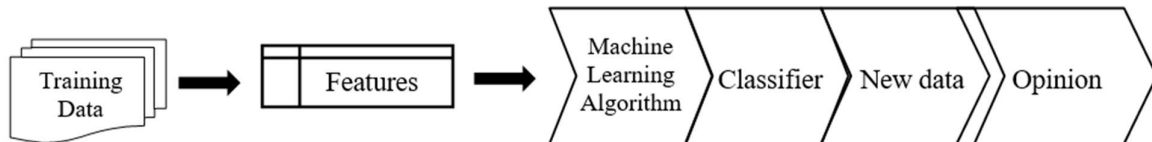


Figure 7. Machine Learning Approach

G. Subjectivity: Detection and Extraction

The automatic detection of the subjectivity target of a sentiment falls into a subarea of information extraction called subjectivity extraction while all the automatic detection of entity's feature, opinion holder and time extraction fall into the more general information extraction. Some specific characteristics of the problem can facilitate the extraction. The key characteristic is that an opinion always has a target. The target is often the aspect or topic to be extracted from a sentence. Thus, it is important to recognize each opinion expression and its target from a sentence. There are some challenges in extracting subjects. Mainly, it is hard to find dictionary to

mine subjects and standardization of subjects. The following figure 8. explains the sample conversion between a patient and health assistant. Table 2. shows the studies Detection and Extraction.

TABLE 2. OVERVIEW OF STUDIES DETECTION AND EXTRACTION

Author	Year	Techniques	Contribution	Conclusion
H Rahman et. al [11]	2022	Multi-layer classification model	In this paper, a multi-tier architecture has been proposed for the multi-class classification of sentiments of the text. The classifiers were trained using three supervised ML algorithms i.e., Naïve Bayes, Support Vector Machine and Decision Tree.	Feature optimization increases the efficiency of the model. Emoticons and POS tagging can be concatenated with the feature vector to deliver true insight into the sentiment.
G Kaur et. al [10]	2023	Combination of Review Related Features (RRF) and Aspect Related Features (ARF)	The input pre-processed reviews can be transformed into meaningful feature vectors, allowing efficient, reliable, and robust sentiment analysis to take place.	The results reveal that as compared to individual methodologies; the hybrid approach greatly improves sentiment analysis performance.
A Mewada et. al [3]	2022	Synthetic attention in bidirectional encoder representations from transformers (SA-BERT) with an extreme gradient boosting (XGBoost)	Generates dynamic word vector encoding of the aspect and corresponding context of the reviews. Then, the aspect and context of the reviews are meaningfully represented by a transformer that can input the vector word in parallel. The model places overall representation in the sentiment classification layer to predict sentiment polarity	Proposed SA-BERT and SA-BERT-XGBoost models achieved the highest accuracy (92.02 and 93.71%)
A Ghosh et. al [1]	2023	Statistical based feature extraction, CNN based end-to-end deep learning framework	In the first component, the face region is detected from the input image. The extracted face region undergoes both statistical and deep learning-based feature analysis in the second component. A prediction model based on statistical feature analysis derives scores; similarly, the deep learning-based prediction model computes scores for the pain intensities (NP, LP, and HP) on the facial region. These scores are used to obtain the performance of the proposed system in the third component. Finally, the scores due to the statistical feature analysis and deep feature analysis are fused to enhance the performance of the proposed method in the fourth component.	This proposed system may be eligible for detecting pain sentiments by analyzing facial expressions on the human face.
Kumar et al. [20]	2018	Supervised learning – Logistic regression, Naive Bayes, CNN	Used for detecting User Emotions in Social Media Text based on logistic regression, Naïve Bayes and CNN	This proposed system has Avg. accuracy: 0.8–0.9 Avg. recall: 0.8–0.9 Avg. precision: 0.8–0.9 Avg. F1: 0.7–0.9.

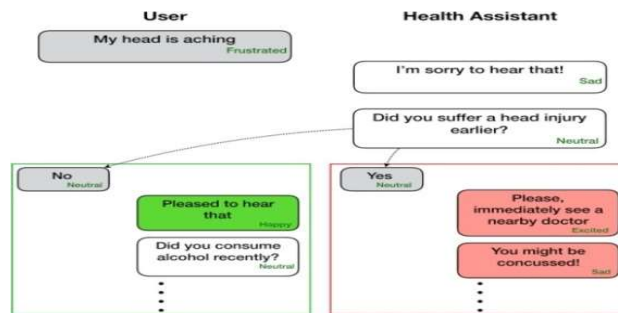


Figure 8. Sample Conversation

H. Degree of Positivity

Sentiment analysis methods are used to identify Sentiment polarity as positive or negative or neutral such as characterizing sentiments into finer subcategories such as anxiety, sadness and anger for negative sentiments or emotions, and excitement and happiness for positive sentiments or emotions as shown in Figure 9.

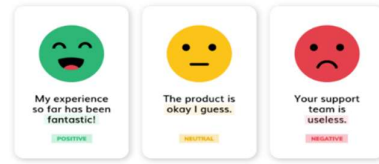


Figure 9. An instance of Positive, Negative and Neutral

JRTrillo et. al [12] proposed Large Scale Group Decision Making method to manage the information generated by a large number of experts, using a sentiment analysis with bag-of-words approach. This approach makes it possible to detect the degree of positivity and aggressiveness of each expert. Once the behaviors have been detected, the experts are grouped according to them and, for each group, a weight and a unique preference relation are obtained. The main advantage is the evaluation of the degree of aggressiveness and positivity of the comments made by the experts.

Z. Wang et. al [26] proposes a new explainable fine-grained multiclass sentiment analysis method, namely MiMuSA, which mimics the human language understanding processes. The proposed method involves a multi-level modular structure designed to mimic human's language understanding processes, e.g., ambivalence handling process, sentiment strength handling process, etc. Specifically, multiple knowledge bases including Basic Knowledge Base, Negation and Special Knowledge Base, Sarcasm Rule and Adversative Knowledge Base, and Sentiment Strength Knowledge Base are built to support the sentiment understanding process. This method understands fine-grained multi-class sentiments, such as the degree of positivity (e.g., strongly positive or slightly positive) and the degree of negativity (e.g., slightly negative or strongly negative) of the sentiments involved.

A M Mostafa [2] provides a comprehensive evaluation of polarity weights and mechanisms for recent sentiment analysis research. A semi-supervised framework is applied for processing data using both lexicon and machine learning algorithms. An interactive sentiment analysis algorithm is proposed for distributing multi-weight polarities on Arabic lexicons that contain high morphological and linguistic terms. An enhanced scaling algorithm is embedded in the multi-weight algorithm to assign recommended weight polarities automatically.

F Benrouba et. al [9] proposing an approach to filter the social media content that could be emotionally harmful to the user, through getting the social networks content; Twitter API to get the user posts and API natural understanding language API tool are used to extract and classify the emotions of the Twitter content into five basic emotional categories—Joy, sadness, anger, fear, disgust—into an array of emotions; after that, a perfect emotion array from over 450 words from the English language was defined. The main purpose of this comprehensive research article is to examine the proposed solution that to improve the quality of content displayed to users emotionally.

I. Topic Modelling

Topic modelling is recognizing the words from the topics present in the document or the corpus of data. The steps of topic modelling depicts in figure 10. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document. Table 3. Gives the review of topic modelling

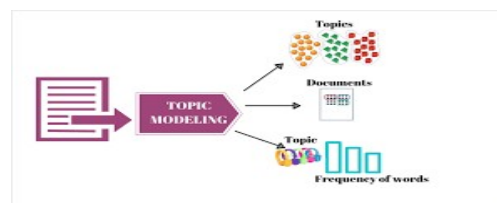


Figure 10. Process of topic modelling

TABLE 3. REVIEW OF TOPIC MODELLING

Author	Year	Techniques	Contribution	Conclusion
D. Sik et al [8]	2021	Quantitative (LDA - topic modelling) and qualitative (deep reading) approaches	13 topics were identified and interpreted according to their content and communicative function. Based on the inter-topic distances four clusters were identified (medicalized, intimacy-oriented, critical and uninhabitable self-narratives)	Expansion of information society and the increased importance of online mental health forums, this research is considered to be a worthy one – not only for theoretical, but also practical reasons.
M. Aleem et. al[14]	2023	Artificial intelligence-based machine learning tools, LDA	Identify-eight research themes: Effect on employees at a personal level, Effect on employees' careers, Family life and gender roles, Health, well-being, and safety, Labor market dynamics, Economic implications, Remote work management, and Organizational remote work strategies	Eight key topics that can be grouped into three broad categories: employee-related research, organization-related research, and labor market and economy-related research. Though the three categories are broadly in line with the extant pre-pandemic research on remote working, the issues within each category have significantly changed.
S. Jiang et. al [22]	2023	Analytical approach, mixed-methods approach	Six major topics about HPV vaccine, namely policy, guidance information, advertising, scandals, personal experience sharing, and HPV risks. Second, an online survey (n = 1,982) to investigate how scanning, seeking, and discussing the six HPV vaccine topics identified from big data analytics can affect HPV vaccine knowledge, safety concern, and vaccination intention.	Different topics or information types exerted dissimilar impacts on health-related outcomes. The same health topic may have differential influences, depending on the information acquisition pattern.
S. Li et. al [21]	2023	Pushshift.io, Reddit API, Regex library	Investigates the sentiments, concerns, and their variance with time regarding online classes during COVID-19 pandemic by learners and educators on Reddit, which is a dominant social network among them. Data were collected via the official API from identified relevant subreddits and keyword search results across Reddit. Sentiment analysis was applied to reveal their emotions and their changes. Topic modeling was conducted to discover the concerns hidden in the posts	Concerns about online classes, such as severe cheating behaviors, and showed doubts about previous strategies to solve disadvantages in online classes. In addition, the results verified the habitual difference and motivations of social media usage between educators and learners.
Wanniarachchi et al.[25]	2023	Mixed analytic approaches	The framework takes hate speech detection a step forward by combining existing NLP techniques to investigate discussion patterns of fat stigma (or hate speech) in a social media context.	The proposed framework is focused on text-based hate speech analysis, while hate speech can be observed in many forms, including visual and audio.

IV. COMPARATIVE STUDY

The various research areas like Sentiment classification, Opinion Identification, Subjectivity detection & Extraction, Degree of positivity and Topic Modelling are explained in above section. Though the direct comparisons cannot be made on performances of the different work as they used different datasets, but looking at the results, we see the reason for the popularity of Sentiment analysis both in machine learning and deep learning various research areas particularly health related issues. The performance of various machine learning techniques are quite similar, thus it can be seen that all methods are used in recent years. The following table 4 depicts the articles' count published in Google Scholar.

TABLE 4. NUMBER OF ARTICLES IN GOOGLE SCHOLAR

Year	Sentiment classification	Opinion Identification	Subjectivity detection & Extraction	Degree of positivity	Topic Modelling
2020	20,300	20,600	16,300	17,400	20,900
2021	22,200	19,700	18,000	18,500	22,600
2022	24,600	22,700	16,500	18,300	22,600
2023	34,600	29,700	16,100	17,900	30,900
2024 (June)	13,000	12,330	11,580	11,460	12,820

V. CONCLUSION

In the above sections, some recent works of sentiment analysis has been summarized to show the methods which are more common in recent years. In the review paper shows machine learning methods on the dataset from large scale patients' narratives, social media and biomedical dataset. In this review, we looked at the recent applications and trends of sentiment analysis which shows that machine learning methods are dominant techniques applied in recent years for health related datasets.

REFERENCES

- [1] A Ghosh et. al Smart sentiment analysis system for pain detection using cutting edge techniques in a smart healthcare framework Cluster Computing (2023) 26:119–135 <https://doi.org/10.1007/s10586-022-03552-z>
- [2] A M Mostafa Enhanced Sentiment Analysis Algorithms for Multi-Weight Polarity Selection on Twitter Dataset Intelligent Automation & Soft Computing (2022) DOI: 10.32604/iasc.2023.028041
- [3] A Mewada et. al SA-ASBA: a hybrid model for aspect-based sentiment analysis using synthetic attention in pre-trained language BERT model with extreme gradient The Journal of Supercomputing (2023) 79:5516–5551 <https://doi.org/>
- [4] Arif Ullah, Sundas Naqeeb Khan & Nazri Mohd Nawawi Review on sentiment analysis for text classification techniques from 2010 to 2021 - Multimedia Tools and Applications (2023) 82:8137–8193 <https://doi.org/10.1007/s11042-022-14112-3>
- [5] Arwa Alqarni and Atta Rahman Arabic Tweets-Based Sentiment Analysis to Investigate the Impact of COVID-19 in KSA: A Deep Learning Approach Big Data and Cognitive Computing (2023) 7, 16. <https://doi.org/10.3390/bdcc7010016>
- [6] Chetashri Bhadane et al Sentiment analysis: Measuring opinions Procedia Computer Science. (2015) 45 808 – 814 www.sciencedirect.com
- [7] D. C. Edara et al. Sentiment Analysis and Text Categorization of Cancer Medical Records with LSTM Journal of Ambient Intelligence and Humanized Computing (2019) 14:5309–5325 <https://doi.org/10.1007/s12652-019-01399-8>
- [8] Domonkos Sik, Renáta Németh & Eszter Katona, Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming, Journal of Mental Health, (2021) DOI: 10.1080/09638237.2021.1979493
- [9] F Benrouba et. al Emotional sentiment analysis of social media content for mental health safety Social Network Analysis and Mining (2023) 13:17 <https://doi.org/10.1007/s13278-022-01000-9>
- [10] G Kaur et. al, A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis Journal of Big Data (2023) <https://doi.org/10.1186/s40537-022-00680-6>
- [11] H Rahman et. al, Multi-Tier Sentiment Analysis of Social Media Text Using Supervised Machine Learning Computers, Materials & Continua, (2022) vol.74, no.3 DOI: 10.32604/cmc.2023.033190
- [12] J.R. Trillo et. al A Large scale Group Decision Making System Based on Sentiment Analysis Cluster SSRN Electronic Journal, (2022) DOI: 10.2139/ssrn.4183223
- [13] Leach, C. W., and Zeineddine, F. B. Sentiments of the dispossessed: Emotions of resilience and resistance. The Routledge international handbook of discrimination, prejudice and stereotyping (2022). (pp. 244–257).
- [14] M. Aleem et. al, Remote work and the COVID-19 pandemic: An artificial intelligence-based topic modeling and a future agenda Journal of Business Research 154 (2023) , 113303
- [15] M. Arbane et al, Social media-based COVID-19 sentiment classification model using Bi-LSTM Expert Systems With Applications 212 (2023) 118710
- [16] Miftahul et. al., Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset, Expert Systems with Applications 212 (2023) 118715

- [17] Nhan et. al, Sentiment Analysis Based on Deep Learning: A Comparative Study Electronics (2020), 9, 483; doi:10.3390/electronics9030483 www.mdpi.com/journal/electronics
- [18] O' Connor B, Balasubramanyan R, Routledge BR, Smith From tweets to polls: linking text sentiment to public opinion time series. (2010) ICWSM 11(122-129):1–2
- [19] Przepiórkowski A XML text interchange format in the National Corpus of polish. In the proceedings of practical applications in language and computers PALC (2009). Peter Lang, Frankfurt
- [20] R. V. Kumar, S. Rahmanian, and H. Al Balooshi, EmotionX SmartDubai NLP: Detecting User Emotions In Social Media Text, Proc. of the 6th Int. Workshop on Natural Language Processing for Social Media, (2018) pp. 45-49, <https://doi.org/10.18653/v1/W18-3508>