A REVIEW OF FEATURE EXTRACTION, REDUCTION AND TRADITIONAL CLASSIFICATION METHODS IN SENTIMENT ANALYSIS

K. Manikandan[1], V. Ganesh[2]

Article History: Received: 08.05.2023 **Revised:** 20.06.2023 **Accepted:** 15.07.2023

Abstract

Sentiment analysis, a crucial component of natural language processing (NLP), aims to extract subjective information from text. This article presents a comprehensive review of feature extraction, reduction, and traditional classification methods in sentiment analysis. The study explores various approaches, including machine learning algorithms, lexicon-based methods, and deep learning models. It discusses the challenges associated with sentiment analysis and highlights the strengths and weaknesses of each technique. The review includes a literature survey of relevant research papers, showcasing innovative methodologies and their application in different domains such as fake news detection, social media sentiment analysis, and public health analysis. The findings emphasize the significance of accurate sentiment analysis in understanding user opinions, public perceptions, and effective communication. The article concludes by discussing the sentiment analysis process and key techniques, including dataset acquisition, pre-processing, feature extraction and reduction methods like Bag of Words (BOW) and TF-IDF, and the use of principal component analysis (PCA) for dimensionality reduction. Further, explores several machine learning algorithms commonly employed in sentiment analysis. These algorithms include Support Vector Machines (SVM), Naive Bayes, Random Forest, and Logistic Regression. SVM is known for its ability to handle high-dimensional feature spaces and non-linear data. Naive Bayes is a probabilistic classifier that assumes feature independence and is efficient for text classification tasks. Random Forest is an ensemble method that combines multiple decision trees to improve classification accuracy. Logistic Regression is a linear model widely used for binary classification tasks. The article discusses the strengths and limitations of each algorithm and highlights their applicability in sentiment analysis tasks. It also presents comparative studies to evaluate the performance of these machine learning algorithms in sentiment classification.

Keywords: Sentiment Analysis, Feature Extraction, Dimensionality Reduction, Machine Learning Algorithms, Performance Evaluation.

[1,2]Department of Computer and Information Science, Faculty of Science, Annamalai University.

DOI: 10.31838/ecb/2023.12.6.187

1. Introduction

Sentiment analysis, also known as opinion mining, is a field of natural language processing (NLP) that involves using computational techniques to identify and extract subjective information from source materials. This can include determining the overall sentiment of a piece of text e.g., positive, negative, or neutral [1], as well as more specific aspects such as identifying the sentiment expressed by individual words or phrases, or determining the sentiment of a particular aspect of the text.

One of the key challenges in sentiment analysis is that the meaning of words can change depending on the context in which they are used. For example, the word "good" might express a positive sentiment in one sentence e.g., "The food was good", but a neutral or negative sentiment in another e.g., "The weather was good for a storm". To address this challenge, many sentiment analysis techniques rely on the use of large amounts of labeled training data, such as product reviews, to "learn" the sentiment associated with different words and phrases.

1.1 Approach

One of the most common approaches to sentiment analysis is the use of machine learning (ML) algorithms. These algorithms can be trained on large datasets of labeled text, and can then be used to classify new text as having a positive, negative, or neutral sentiment. A variety of different types of ML algorithms have been used for sentiment analysis, including decision trees, Naive Bayes, and support vector machines (SVMs) [2].

Another popular approach to sentiment analysis is the use of lexicon-based methods [1], which rely on lists of words and their associated sentiment scores. For example, a lexicon-based method might use a list of positive and negative words, and assign a positive or negative sentiment score to a piece of text based on the number of positive or negative words it contains. These methods can be simple and fast, but they can also be less accurate than ML-based methods, particularly in cases where the sentiment of a word is dependent on the context in which it is used.

1.2 Challenges

Sentiment analysis is a challenging field of NLP [3] that involves using computational techniques to identify and extract subjective information from text. A variety of different techniques have been developed to address this challenge, including machine learning, lexicon-based methods, and deep learning models. Each of these techniques has its own strengths and weaknesses, and the best approach will depend on the specific task and dataset at hand.

2. Literature Survey

Yuxing Qi and Zahratu Shabrina [1] The literature highlights the significance of social media platforms, like Twitter, as a virtual arena for users to share their thoughts, opinions, and engage in interactions through succinct 140-character posts called tweets. These tweets can take various forms, encompassing texts, pictures, videos, and more, while user interaction is facilitated through features like likes, comments, and reposts. With the exponential growth in social media participation, analyzing the wealth of online information has become instrumental in understanding shifts in people's perceptions, behavior, and psychology. Consequently, leveraging Twitter data for sentiment analysis has emerged as a prominent trend in research. This article specifically aims to compare the performance of lexicon-based and machine learning-based approaches for sentiment analysis using Twitter data. The

lexicon-based approach employs a sentiment lexicon to assign sentiment scores to individual words, whereas the machine learning-based approach utilizes machine learning algorithms to discern the sentiment expressed in tweets. The findings of this study demonstrate that the machine learning-based approach surpasses the lexicon-based approach in terms of accuracy. However, it is acknowledged that the lexicon-based approach offers advantages in terms of implementation speed and ease. Ultimately, the choice of approach is contingent upon the specific application at hand. If accuracy holds paramount importance, researchers are advised to opt for the machine learning-based approach. Conversely, if speed and ease of implementation are prioritized, the lexicon-based approach proves to be a favorable option. Additionally, the paper delves into the challenges inherent in sentiment analysis of Twitter data, such as the utilization of slang, emojis, and abbreviations. Moreover, the article presents potential methods to address these challenges, offering valuable insights for future research in this domain.

Sarita V Balshetwar et al [3] an innovative approach is presented in this study, aiming to enhance the accuracy of fake news detection by incorporating sentiment as a significant feature. The proposed solution was evaluated using two distinct datasets, namely ISOT and LIAR, to ensure a comprehensive analysis and validation. To develop the key feature words that capture the content's propensity scores of opinions, the study employed sentiment analysis through a lexicon-based scoring algorithm. This approach facilitated the identification of crucial words that contribute to the overall sentiment expressed in the text. Furthermore, the study addressed the issue of missing variables in social media or news data by proposing a multiple imputation strategy that integrated the Multiple Imputation Chain Equation (MICE) technique. By leveraging MICE, the researchers effectively handled multivariate missing variables within the collected dataset. Additionally, the study introduced the utilization of Term Frequency and Inverse Document Frequency (TF-IDF) to extract effective features from the textual data. This technique enabled the determination of long-term features by constructing a weighted matrix. The correlation between missing data variables and useful data features was then classified using Naïve Bayes, passive-aggressive, and Deep Neural Network (DNN) classifiers. The findings of this research underscored the significant achievements of the proposed method in the detection of fake news. With an impressive accuracy rate of 99.8%, the proposed method successfully identified fake news across various statements, including barely true, half true, true, mostly true, and false, within the dataset. These results demonstrate the effectiveness of the proposed method in accurately categorizing and detecting fake news instances. Furthermore, to gauge the performance of the proposed method, a comparative analysis was conducted with existing methods. The results of this comparison showcased the superior efficiency of the proposed method, highlighting its potential as a robust solution for fake news detection. By presenting these advancements and breakthroughs, this study contributes to the existing literature on fake news detection. The incorporation of sentiment as an important feature, coupled with the development of key feature words, the application of multiple imputation strategies, and the utilization of advanced classifiers, reinforces the accuracy and reliability of the proposed method. The findings of this research provide valuable insights into the evolving field of fake news detection, paving the way for improved techniques and approaches to combat misinformation and promote a more trustworthy information ecosystem.

Murtuza Shahzad et al [4] Social media platforms offer diverse means of user interaction, such as commenting, reacting to posts, sharing content, and uploading pictures. Among these platforms, Facebook stands out as one of the most popular, where users frequently engage in sharing and resharing posts, including those pertaining to research articles. Notably, Facebook's reactions feature enables users to express their sentiments towards the content they encounter, thereby presenting valuable data for analysis. This particular research endeavors to forecast the emotional impact of Facebook posts associated with research articles. To accomplish this, we collected data from Facebook posts encompassing various scientific research domains, including Health Sciences, Social Sciences, Dentistry, Arts, and Humanities. Our investigation revolved around examining Facebook users' reactions to both research articles and accompanying posts. Remarkably, we found that 'Like' reactions were the most prevalent among users. Additionally, we observed that research articles within the Dentistry domain garnered a considerable number of 'Haha' reactions. In order to predict the sentiment of Facebook posts associated with research articles, we employed machine learning models. These models utilized features such as the sentiment of the research article's title, abstract sentiment, abstract length, author count, and research domain. We tested five classifiers, namely Random Forest, Decision Tree, K-Nearest Neighbors, Logistic Regression, and Naïve Bayes, and evaluated their performance using accuracy, precision, recall, and F-1 score metrics. Among these classifiers, the Random Forest model emerged as the most effective for two- and three-class labels, achieving accuracy measures of 86% and 66%, respectively. Additionally, we conducted a feature importance analysis for the Random Forest model, revealing the sentiment of the research article's title as a critical factor in predicting the sentiment of the corresponding Facebook post. This study holds significant implications for public engagement with science-related messages. The emotional reactions exhibited by Facebook users towards research articles and posts furnish valuable insights into public engagement with scientific endeavors. Moreover, the ability to forecast the emotional impact of Facebook posts linked to research articles aids researchers in comprehending how the public perceives scientific research. Consequently, the findings of this study can facilitate effective communication of research findings and foster public engagement in scientific discourse.

Staphord Bengesi et al [8] Sentiment analysis research has proven invaluable in public health, specifically in the analysis of infectious diseases. With the world recovering from the devastating impact of the COVID-19 pandemic, concerns are growing about the potential resurgence of another infectious disease called monkeypox. Monkeypox has been reported in over 73 countries worldwide, leading to increased worries among individuals and health authorities. Social media platforms have become channels for discussions, opinions, and emotions surrounding the monkeypox outbreak. However, these sentiments often give rise to panic, misinformation, and stigmatization of marginalized groups. Thus, the availability of accurate information, guidelines, and health protocols related to this virus is crucial. This study aims to analyze the public sentiments regarding the recent monkeypox outbreak, providing decision-makers with a better understanding of public perceptions. The findings can aid government and health authorities in formulating effective health policies, implementing mitigation strategies, and combating the spread of the disease while addressing misrepresentations. The study was conducted in two stages. Initially, over 500,000 multilingual tweets related to

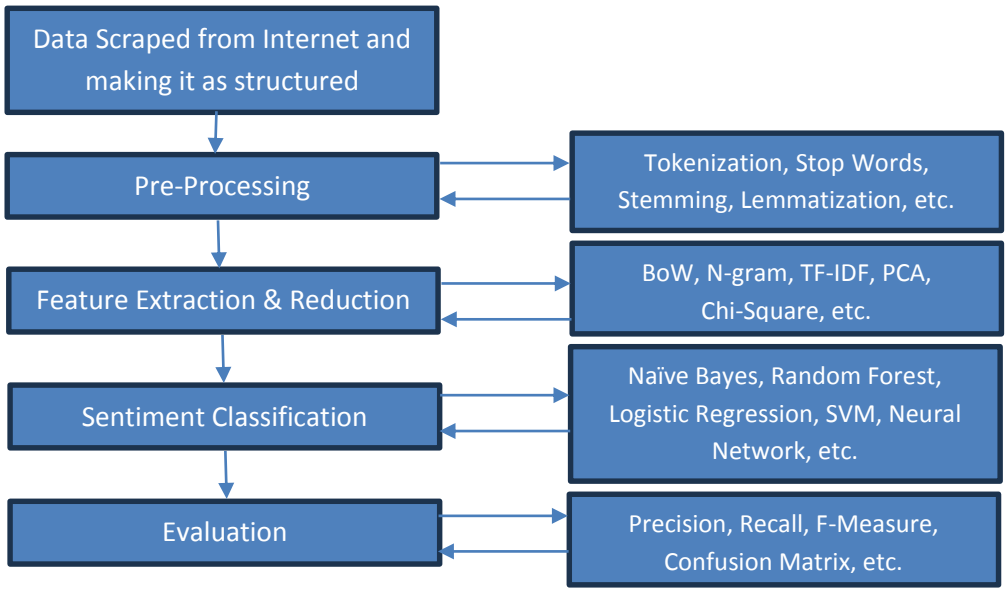
monkeypox were collected from Twitter, and sentiment analysis was performed using VADER and TextBlob to annotate the tweets as positive, negative, or neutral sentiments. The second stage involved the design, development, and evaluation of 56 classification models. Vocabulary normalization was achieved through stemming and lemmatization techniques. CountVectorizer and TF-IDF methodologies were employed for vectorization. Learning algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Logistic Regression, Multilayer Perceptron (MLP), Naïve Bayes, and XGBoost were utilized. Performance evaluation was based on accuracy, F1 Score, Precision, and Recall. Experimental results revealed that the model constructed using TextBlob annotation, Lemmatization, CountVectorizer, and SVM achieved the highest accuracy of approximately 0.9348.

3. Methodology

Sentiment analysis performed on different levels: sentence level, document level and aspect level [7]. Sentiment analysis can be stated as the procedure to identify, recognize, and/or categorize the users' emotions or opinions for any service like movies, product issues, events, or any attribute as positive, negative, or neutral. When sentiment is stated as a polarity in computational linguistics, it is typically treated as a classification task. When sentiment scores lying inside a particular range are used to express the emotion, the task is however regarded as a regression problem. Sentiment analysis is approached as either a classification or regression task, while analyzing the sentiments by assigning the instances sentiment scores within the range $[-1,1]$, there can be circumstances where the prediction is sometimes considered to be a classification task and other times to be regression. Mining and analysis of sentiment are either limited to positive/negative/neutral; or even deeper granular sentimental scale, depending on the necessity, topic, scenario, or application.

4. Sentiment Analysis Process

Sentiment Analysis Framework



4.1 Dataset

Scraping data from social networking websites, reviews from product websites, e-commerce websites, forums, weblog, etc. can be used for analysis [5]. Since it's a classification process, the algorithm needs to be trained. So, the training dataset makes the developed model more efficient in predicting the output. In-order to evaluate the model test dataset is used.

4.2 Pre-Processing

Pre-Processing [6] is the process of cleaning and preparing the text for sentiment classification, which includes removing URLs, removing numbers, removing repeated letters, removing stop words; stemming and expanding acronyms are applied to reduce the amount of noise in tweets. Many words in the corpus do not have direct impact on the general orientation on it. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension. Based on the nature of dataset or domain the pre-processing methods are applied before feature extraction. Each step in the pre-processing affects the performance of the classifier algorithm and speed up the classification process.

4.3 Feature Extraction and Reduction Methods

BOW (Bag of Words) [6] also known as Vector Space Model (VSM) is an algebraic model used to represent text as vectors of documents. The dimensionality of the vector is the number of words in the vocabulary i.e., the number of distinct words occurred in the corpus. BOW represents the document term matrix i.e., document, vocabulary and occurrence.

N-Gram [6] is a method of 'n' continuous words. Unigram refers to n-gram of size 'one,' Bigram refers to the n-gram of size 'two,' and Trigram refers to the n-gram of size 'three' and so on. The combination of n-gram can also be used.

Term Frequency and Inverse Document Frequency (TF-IDF) [6-8] used for feature weighting; it is a standard approach to feature vector construction. TF-IDF stands for the "term frequency-inverse document frequency" and is a numerical statistic that reflects how important a word is to a document in a corpus.

$$TF \cdot IDF = TF_{i,j} \times IDF_{i,j} = TF_{i,j} \times \log(N /DF_j)$$

N = number of documents in collections.
TF = term frequency.
IDF = inverse document frequency.

Principal Component Analysis (PCA)

PCA [6] based feature reduction process takes place to get the reduced set of features. It belongs to unsupervised linear conversion model which is applied over diverse applications, for feature extraction as well as dimensionality minimization. PCA assist in identifying patterns from a data which depends upon correlation among features. The main goal of PCA is to explore the directions of higher variance in high-dimensional data and present in a novel subspace with same dimensions. The orthogonal axes of new subspace could be disturbed as directions of maximum variance provided with a constraint which has novel axes are said to be orthogonal with one another as depicted in the given Fig. 1. In this figure, x1 and x2 are represented by actual feature axes, PC1 and PC2 are named as principal components.

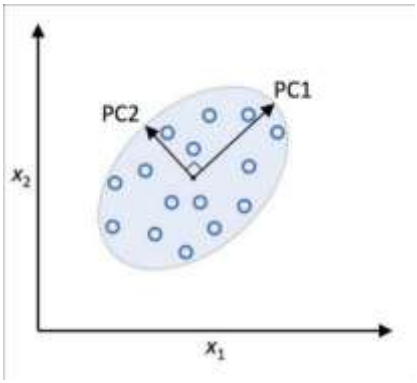


Fig. 1. Orthogonal axes on PCA

In case of PCA for dimensionality reduction, it is developed with d x k–dimensional transformation matrix W which enables to map a sample vector x onto a novel k–dimensional feature subspace. The actual d-dimensional data is transformed as novel k-dimensional subspace. The primary PCA has maximum feasible variance and every final principal component might consist of largest variance, provided the constraint is uncorrelated with another PCA would be mutually orthogonal. The process involved in PCA based feature reduction process is given below.

- Standardize the d-dimensional dataset.
- Develop the covariance matrix.
- Degraded the covariance matrix into its eigenvectors and Eigen values.
- Types of Eigen values by decreasing order to range the corresponding eigenvectors.
- Choose k eigenvectors which correspond to k largest Eigen values, where k denotes dimensionality of new feature subspace ($k \leq d$).
- Develop a projection matrix W from "top" k eigenvectors.
- Convert the d-dimensional input dataset X with the help of matrix W to attain the novel k-dimensional feature subspace.

Chi-Square

Chi-Square [6] is use to measure the association between terms and class label or categories. Let n be the total number of documents in the collection, Pi (w) be the conditional probability of class i for documents which contain w, F(w) be the global fraction of documents which contain the word w, and Pi be the global fraction of documents containing the class i. Therefore, the v2-statistic of the word between word w and class i is defined as

$$x_i^2 = \frac{n.F(w)^2.(P_i(w) - P_i)^2}{F(w).\big(1 - F(w)\big).P_i.(1 - P_i)}$$

4.4 Traditional Machine Learning Algorithms in Sentiment Classification

Navie Bayes

The Naïve Bayes [9] was an algorithm adopted for classification. It is a probabilistic classifier that uses conditional probability to determine the class likelihood of its input. The Naive Bayes is the simplest statistical and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. The particular label is predicted by using Bayes Theorem with given feature set.

$$P(\text{label} \mid \text{features}) = \frac{P(\text{label}) * P(\text{features} \mid \text{label})}{P(\text{features})}$$

P (label) is the prior probability of a label or the likelihood that a random feature set the label. P (features | label) is the prior probability that a given feature set is being classified as a label. P (features) is the prior probability that a given feature set is occurred. Given the Naive assumption which states that all features are independent, the equation could be rewritten as follows

$$P(l \mid f) = \frac{P(l) * P(f1 \mid l) * \ldots * P(fn \mid l)}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

l—Label, f —Features.

Random Forest

Random Forest [10] is another model we utilized. It is an ensemble machine-learning classification algorithm. This algorithm develops numerous decision trees used for classification, which implies that class category is selected by most of the trees. This approach involves randomization and aggregation of tree prediction into the final output. Random forest requires at least three hyperparameters to be in place: node size, number of trees, and number of features sampled. It applies bootstrap aggregation, also known as the bagging ensemble technique, which creates a different subset of training adopted from sample training data. The result depends on the rate of preference.

Logistic Regression

Logistic regression [12] can be defined as the supervised learning algorithm which predicts the probability of an event occurrence. It is a statistical modeling technique used to predict binary outcomes by estimating the probability of an event occurring. It is commonly employed when the dependent variable is categorical, with two possible outcomes (e.g., yes or no, pass or fail). The logistic regression model uses the logistic function to transform a linear combination of predictor variables into a probability value between 0 and 1. The formula for logistic regression can be expressed as:

p = 1 / (1 + e^(-z))

Where:

p represents the probability of the event occurring,

e is the base of the natural logarithm (approximately 2.718),

z is the linear combination of predictor variables and their coefficients:

$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta\square x\square$

In the equation, $\beta_0, \beta_1, \beta_2, \ldots, \beta\square$ are the coefficients associated with the predictor variables $x_1, x_2, \ldots, x\square$. These coefficients are estimated using maximum likelihood estimation to optimize the model's fit to the data. The logistic function transforms the linear combination of predictor variables into a probability, which can then be thresholded to make binary predictions based on a specified cutoff value.

Support Vector Machine (SVM)

SVM [11,13] is a robust model that sets boundaries between classes, sorting data into one of the available categories. SVM needs decision boundaries (separator lines) between classes called a hyperplane. There exist three hyperplanes, namely positive, negative, and optimal hyperplanes.

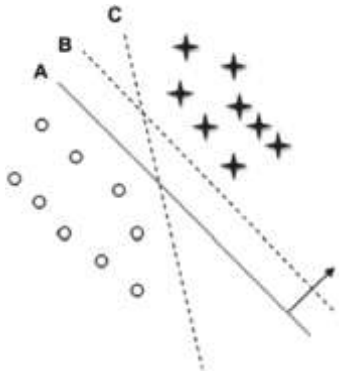


Fig. 2. Support vector machine—Linear separator with two classes x, o.

Text data are preferably suited for SVM classification due to the sparse nature of text, in which few features are inappropriate, but they tend to be correlated with one another and generally organized into linearly separable categories. SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyperplane.

4.5 Evaluation

To measure the correctness of classifier precision is calculated [16]. While recall measures the completeness or sensitivity of a classifier. The f -measure [16] is the weighted harmonic mean of precision and recall. The accuracy [16] will evaluate how often the classifier is correct.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$\text{F} - \text{Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

5. Conclusion and Future Work

In conclusion, this survey paper provides a comprehensive understanding of how sentiment classification challenges are addressed using machine learning techniques. The utilization of diverse pre-processing methods has significantly enhanced the accuracy of specific algorithms [19]. To evaluate the effectiveness of the proposed methods and algorithms, benchmark datasets from various domains have been employed. The results are compared and visually presented to highlight the performance. Moving forward, future research can focus on incorporating advanced techniques such as neural networks [17,18] and deep learning, instead of relying solely on traditional machine learning algorithms. Furthermore, adapting feature techniques based on the specific nature of the domain can further enhance accuracy.

6. References

1. Yuxing Qi, Zahratu Shabrina (2023) Sentiment analysis using Twitter data: a comparative application of lexicon and machine learning based

approach, Social Network Analysis and Mining (SPRINGER), https://doi.org/10.1007/s13278-023-01030-x

2. Sarita V Balshetwar, Abilash RS, Dani Jermisha R (2023) Fake news detection in social media based on sentiment analysis using classifier techniques. Multimedia Tools and Applications (SPRINGER). https://doi.org/10.1007/s11042-023-14883-3.
3. Sayar Ul Hassan, Jameel Ahamed, Khaleel Ahmad (2022) Analytics of machine learning-based algorithms for text classification, Sustainable Operations and Computers, 238–248
4. Murtuza Shahzad, Cole Freeman, Mona Rahimi, Hamed Alhoori (2023) Predicting Facebook sentiments towards research, Natural Language Processing Journal (ELSEVIER). https://doi.org/10.1016/j.nlp.2023.100010
5. Prasoon Gupta, Sanjay Kumar, R. R. Suman, and Vinay Kumar (2021) Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter, IEEE Transactions on Computational Social Systems.
6. V. Ganesh, M. Kamarasan (2019), Sentiment Classification and Its Feature Models: A Survey, Journal of Computational and Theoretical Nanoscience (Vol. 16, 1–5), doi:10.1166/jctn.2019.8068
7. Priyavrat Chauhan, Nonita Sharma, Geeta Sikka (2020) The emergence of social media data and sentiment analysis in election prediction, Journal of Ambient Intelligence and Humanized Computing (SPRINGER), https://doi.org/10.1007/s12652-020-02423-y
8. Staphord Bengesi, Timothy Oladunni, Ruth Olusegun, Halima Audu (2023) A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion from Twitter Tweets, IEEE Open Access Journal, DOI: 10.1109/ACCESS.2023.3242290
9. Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath (2016) Classification of sentiment reviews using n-gram machine learning approach, Expert Systems with Applications (ELSEVIER), http://dx.doi.org/10.1016/j.eswa.2016.03.028
10. Zhao Jianqiang, Gui Xiaolin (2017) Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, IEEE Open Access Journal, DOI: 10.1109/ACCESS.2017.2672677
11. Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, Linda Galligan (2023), Sentiment analysis and opinion mining on educational data: A survey, Natural Language Processing Journal (ELSEVIER), https://doi.org/10.1016/j.nlp.2022.100003
12. Pardeep Kaur, Maryam Edalati (2022) Sentiment analysis on electricity twitter posts, Computation and Language, arXiv:2206.05042v1
13. Hermawan Syahputra, Aldiva Wibowo (2023) Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites, Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), DOI: 10.26555/jiteki.v9i1.25861
14. Raghda Elnadrey, Ashrif Elsisi, Walid Attwa (2022) Performance Investigation of Features Extraction and Classification Approaches for Sentiment Analysis Systems, International Journal of Computers and Information (IJCI), DOI: 10.21608/IJCI.2021.65578.1044

15. Hassan Nazeer Chaudhry, Yasir Javed, Farzana Kulsoom, Zahid Mehmood, Zafar Iqbal Khan, Umar Shoaib, Sadaf Hussain Janjua (2021) Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020, Electronics (MDPI), https://doi.org/10.3390/electronics10172082
16. A. Naresh, P. Venkata Krishna (2020) An efficient approach for sentiment analysis using machine learning algorithm, Evolutionary Intelligence (SPRINGER), https://doi.org/10.1007/s12065-020-00429-1
17. Pansy Nandwani, Rupali Verma (2021), A review on sentiment analysis and emotion detection from text, Social Network Analysis and Mining (SPRINGER), https://doi.org/10.1007/s13278-021-00776-6
18. Sofia, Arun Malik, Mohammad Shabaz, Evans Asenso (2023), Machine learning based model for detecting depression during Covid-19 crisis, Scientific African (SPRINGER), https://doi.org/10.1016/j.sciaf.2023.e01716
19. Nasir Jalal, Arif Mehmood, Gyu Sang Choi, Imran Ashraf (2022) A novel improved random forest for text classification using feature ranking and optimal number of trees, Journal of King Saud University – Computer and Information Sciences (SPRINGER), https://doi.org/10.1016/j.jksuci.2022.03.012