Journal of Informatics and Web Engineering

Vol. 3 No. 1 (February 2024) eISSN: 2821-370X

Sentiment Analysis using Support Vector Machine and Random Forest

Talha Ahmed Khan[1*], Rehan Sadiq[2], Zeeshan Shahid[3], Muhammad Mansoor Alam[4], Mazliham Bin Mohd Su'ud[5]

[1, 5] Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.
[2] Computer Sciences, Bahria University Karachi Campus, 13 National Stadium Rd, Karsaz Faisal Cantonment, Karachi, Pakistan.
[3]Faculty of Engineering, Nazeer Hussain University, ST-2, Block-4, Federal B Area, Near Karimabad, Karachi, Pakistan.
[4]Faculty of Computing, Riphah International University, Karachi, Pakistan
[*]*Corresponding author: (ahmedkhan.talha01@mmu.edu.my; ORCiD: 0000-0001-6687-0920)*

Abstract - Sentiment analysis, is commonly known as opinion mining, is a vital field in natural language processing (NLP) that claims to find out the sentiment or emotion expressed in a given text. This research paper demonstrates an exhaustive survey of sentiment analysis, focusing on the application of machine learning techniques. Comprehensive parametric literature review has been completed to determine the sentiment analysis using SVM and Random Forest. Additionally, the paper covers preprocessing techniques, feature extraction, model training, evaluation, and challenges encountered in sentiment analysis. The findings of this research contribute to a deeper understanding of sentiment analysis and provide insights into the effectiveness of machine learning approaches in this domain. Based on the results obtained, two machine learning algorithms named as Random Forest and SVM were evaluated based on their accuracy in a classification task. The Random Forest algorithm achieved an accuracy of 0.78564, while SVM outperformed it slightly with an accuracy of 0.80394. Both Random Forest and SVM have demonstrated their strengths in achieving respectable accuracies in the given classification task. These results suggest that SVM, with its slightly higher accuracy of 0.80394, may be a more suitable choice when accuracy is the primary concern. However, the basic configuration need and characteristics of the problem at hand should be considered when choosing the better algorithm with better results.

Keywords— Sentiment Analysis, Machine Learning, Opinion Mining, Natural Language Processing, Preprocessing Techniques, Feature Extraction

Received: 7 June 2023; Accepted: 3 September 2023; Published: 16 February 2024

I. INTRODUCTION

The paper begins with an introduction to sentiment analysis, followed by a discussion of various machine learning algorithms and methodologies employed in sentiment analysis tasks. In recent years, the Internet has experienced significant expansion, offering a convenient platform for individuals to share their opinions. With the accessibility and anonymity it provides, people are increasingly inclined to express their viewpoints on social media platforms like Facebook and twitter. This has led to a diverse and multifaceted expression of opinions on various social topics. As a result, sentiment analysis and evaluations, aimed at understanding the prevailing sentiment in comments on platforms like Facebook and twitter, have emerged as a famous research area within the academic community. Sentiment analysis, a significant field in Natural Language Processing (NLP) utilizing NLP and text mining techniques, involves analyzing and extracting the emotional tendencies embedded in subjective text. It has garnered extensive research attention. By evaluating comments on social media, sentiment analysis enables the

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2023.3.1.5

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

determination of whether expressions are positive or negative. The extraction of sentiment and emotional insights from comments has made sentiment analysis widely adopted and valuable.

II. SENTIMENT ANALYSIS: OVERVIEW AND CHALLENGES

Sentiment analysis which is commonly known as opinion mining, involves identifying and extracting subjective based information from text data [1]. It plays a crucial role in understanding public sentiment, customer feedback analysis, brand monitoring, and market research [2]. Sentiment analysis can be performed with various techniques, including machine learning, lexicon-based methods, and hybrid approaches that combine multiple techniques [3], Sentiment analysis face hurdles and challenges such as sarcasm, irony, context-dependent sentiment, and handling noisy or imbalanced datasets. Overcoming these challenges is essential for accurate sentiment classification [4].

III. PRE-PROCESSING TECHNIQUES

Figure 1 explains tokenization involves splitting text into individual words, phrases, or sentences to facilitate further analysis. Common techniques include whitespace-based tokenization, rule-based tokenization, and statistical models such as the Penn Treebank tokenizer [5], Stop words are commonly occurring words (e.g., "the," "and," "is") that do not carry significant meaning in sentiment analysis. Removing stop words can reduce noise and improve computational efficiency [6], Stemming reduces words to their root form (e.g., "running" to "run"), while lemmatization maps words to their base or dictionary form (e.g., "better" to "good"). These techniques help in normalizing text and reducing feature space [7]. Negations and emoticons can affect sentiment polarity. Techniques such as "not" handling and sentiment-specific emoticon mapping can improve sentiment analysis accuracy [8].

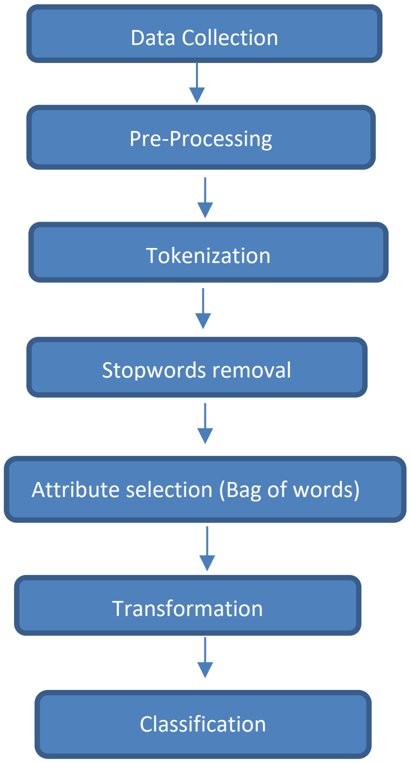

Figure 1: Fundamentals of Pre-processing Techniques

Special characters, URLs, and HTML tags need to be properly handled or removed during preprocessing to avoid interference with sentiment analysis algorithms. Regular expressions and specialized libraries can be utilized for this purpose [9], converting all text to lowercase can help in reducing the dimensionality of the feature space and ensuring consistency in sentiment analysis [10], Spelling errors can affect sentiment analysis results. Spell checking and correction techniques, such as using language-specific dictionaries or algorithms like Levenshtein distance, can be employed to address this issue [11]. The Bag of words (BoW) model represents text as a collection of words and their frequencies, disregarding word order. It forms a matrix of word occurrences, which can be used as input for machine learning algorithms [12].

Term frequency and inverse document frequency (TF-IDF) measures the importance of a word in a document corpus by considering both term frequency (how often a word appears in a document) and inverse document frequency (how common or rare a word is across the entire corpus) [13]. Word embedding capture semantic relationships between words by representing them as dense vectors in a continuous vector space. Techniques such as Word2Vec and GloVe can generate word embedding that capture contextual information [14].

IV. EXISTING APPROACHES

The concept of sentiment analysis, originally proposed by Rosalind W. Picard [14], involves the calculation and analysis of moods, feelings, and opinions in text. It enables the determination of emotional tendencies within subjective content. Presently, this technology predominantly focuses on analyzing comments from internet users concerning products, services, and current events. These texts provide valuable insights into the genuine emotional inclinations of internet users, serving as a reference for other potential users. Scholars continuously refine algorithms and apply sentiment analysis techniques across various fields. Sentiment analysis techniques have been applied in diverse domains.

Alfarrarjeh et al. [15] examine disaster-related social media posts, Maia et al. [16] propose a sentiment-based prediction model for financial decision-making, Vanaja et al. [17] analyze E-commerce data to improve customer understanding and sales, and Anu [18] conducts a comparative study of Twitter sentiment on Covid-19 tweets, offering insights for government and health officials. Sentiment analysis technology research methods can be broadly categorized into sentiment dictionary-based and machine learning-based approaches. Emotion dictionaries form the foundation of sentiment analysis. Whisell et al [19] enhance the sentiment dictionary for improved natural language compatibility, while Seongik Park et al. [20] expand and enhance the dictionary's usability by recursively collecting synonyms and antonyms. Sentiment dictionaries find wide application in public opinion feature extraction, comment analysis, and cross-domain sentiment classification. Machine learning-based sentiment analysis methods can be categorized into supervised, semi-supervised, and unsupervised approaches.

Pang et al. [21] conducted a comparative study of naive Bayes, maximum entropy classification, and support vector machine for classifying film review emotions. Hemalatha et al. [22] employed naive Bayes and other algorithms to analyze sentiment in Yelp reviews. Supervised learning is widely adopted and continually advancing. In contrast, unsupervised learning does not require manual labeling but sacrifices accuracy. Sun et al. [23] proposed an unsupervised topic emotion model that combines unsupervised machine learning with LDA topic modeling. Semi-supervised learning, which leverages a small amount of labeled data, has gained popularity. Jin et al. [24] developed a semi-supervised learning model by integrating the TRS-SAT model with a convolutional neural network. Scholars employ diverse methodologies for analyzing social media comments. Thao Thanh Nguyen [25] utilizes the Facebook Graph API to identify appealing fields based on Facebook comments. Anna A. Gamova [26] develops a deep learning system for identifying and categorizing negative online content, enhancing the online experience [27].

Scholars analyze sentiment in social media comments to apply the findings across different domains. R. Meena [28] demonstrates that individuals rely more on tweets shared by professionals in health-related Twitter analysis [29]. This model utilizes a dataset of 100,000 open-source Weibo comments labeled with emotion categories The data is split into a training set and a test set (8:2 ratio) for model development and validation. By comparing random forest and support vector machine algorithms, support vector machine is chosen as the foundation for the model algorithm [30-34]. NER rule based approach depends on defined boundaries, these techniques help to achieve efficiency and better accuracy rate but cannot be used with other domains [35]. The dataset used for this research belonged to the IBM search panel including many queries.

The researcher mentioned the same dataset used for both training and testing and applied to all models. The writer also focused on the accuracy F1-score, Time-prediction and for the training. For the future, researchers mentioned that the semantic approach can be used while applying NER technique. It is also stated the biggest challenge is to find the dataset for low resource language [36-37].

Current automated log analytics systems in the modern era do not take variables into account when processing log events. However, variables, such as the return code (e.g., "404") found in logs, hold significant meaning in terms of the system's operational status. To address the critical obstacle of extracting these semantic meanings from log messages, this study introduces LogVM, which consists of three components: (1) an encoder that captures contextual information, (2) a pair matcher that resolves the semantics of variables, and (3) a word scorer that disambiguates different semantic roles. Through experiments conducted on seven widely-used software systems, LogVM successfully extracts comprehensive semantics from log messages. It was considered that uncovering these variable semantics can greatly support downstream applications for system maintainers [38]. A lot of challenges and hurdles are faced by the search engines due to the spam web pages. The spam content is commonly being utilized over the network. The spam content is very complex to be detected as the internet technologies have been developed a lot. The existing approach for the spam detection in web pages is primarily based on the statistical features which has dominantly limitations. In this article two spam detection methods have been adopted to determine the spam in the web page named as semantics and statistics. Web page contents have been mapped into the topic space and the modeling for the topic was performed on the web page contents. Distribution of the topic was calculated and it was followed by the semantic analysis. Statistical analysis with the combination of feature extraction was adopted for the classification of spam web pages to determine the sematic features. Results proved that the proposed approach could achieve better results [39].

The model's efficacy is further demonstrated through sentiment classification and cluster analysis of event review data across different categories (positive/negative), confirming its soundness. The data undergoes preprocessing tasks like word segmentation and text vectorization before fitting the algorithm [40]. The study compares the random forest and support vector machine algorithms for model selection. In the random forest approach, data and features are sampled with replacement, ensuring that each decision tree is trained on a subset that maintains a balanced scale. Predictions are made by aggregating the results from all decision trees through voting, effectively preventing over fitting. On the other hand, support vector machine builds the model by so1ving for an optimal separation boundary, maximizing the distance between the closest points and the resulting hyper plane, enabling binary classification. Following the partitioning of the dataset into training and test sets (8:2 ratio), the support vector machine and random forest algorithms are compared in terms of accuracy. Based on the results, this study selects the support vector machine algorithm for model construction. Once the model is established, the study focuses on selecting social hot events characterized by intense emotional tendencies. Leveraging the constructed model, sentiment classification is performed on social media comments provided by users regarding these events. The approach employed in this paper involves binary classification, categorizing social media comments into positive and negative emotion categories. After sentiment classification, the paper employs the DBSCAN algorithm for cluster analysis. DBSCAN is a density-based clustering method that can identify clusters of arbitrary shapes and does not require predefining the number of clusters. It helps reveal valuable emotion categories and prevailing viewpoints [41].

V. IMPLEMENTATION

While it's a fact that news articles can sometimes contain inaccuracies, research doesn't hinge solely on the absolute accuracy of every individual piece of data. Instead, it revolves around the identification of keywords and their contextual meaning within the occurrences of these keywords. To ensure the reliability of the data. It was manually compiled, and taken the step of crawling technology articles from reputable and trustworthy news websites spanning over a decade. While manual selection does have its inherent limitations, my dataset is backed by authentic sources, strict criteria for object selection, and a transparent process that clearly outlines all the steps taken in its manual compilation.

The evaluation of the model's performance in this study involves cross-validation scores on the training set and comprehensive metrics on the test set. By utilizing cross-validation, over fitting is mitigated, and the model's fit and generalization ability are assessed effectively. When using supervised learning algorithms like decision trees or random forests, using all data for training can lead to high scores but poor generalization. To address this, cross-validation is employed to evaluate the model's performance by dividing the data into subsets for training and validation. This method comprehensively assesses the model's fit and generalization ability. Additionally, metrics such as accuracy, recall, precision, and F1-score are used to evaluate the model on unseen test data. Accuracy represents the model's classification accuracy by measuring the ratio of correct predictions to total predictions. Accuracy is a metric that measures the classification precision of a model by calculating the proportion of correct predictions out of the total number of predictions made. It reflects the model's ability to accurately classify instances.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

When negative outcomes make up a significant portion of the data and positive outcomes are relatively rare, a sub-model can achieve high accuracy by simply predicting all instances as negative. However, such a model lacks predictive capability. To address this, the recall metric becomes essential. Recall measures the proportion of positive predictions to the actual positive outcomes. In the aforementioned scenario, even with a high accuracy, the recall would be 0 since the model achieves high accuracy solely through negative predictions, thus missing all positive predictions. TP is acknowledged as true positive and TN is acknowledged as true negative. FP means false positive and while FN represents the false negative. These values have been estimated using support vector machine and random foresee classifier confusion matrix.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

When negative results outweigh positive results, predicting all instances as positive may result in low accuracy but high recall. In such cases, precision tends to be low, representing the ratio of correct positive predictions to all positive predictions. TP is acknowledged as true positive and TN is acknowledged as true negative. FP means false positive and while FN represents the false negative. These values have been estimated using support vector machine and random foresee classifier confusion matrix.

A. Text Vectorization

Term frequency and inverse document frequency (TF-IDF) technique was applied to convert text into numerical representations. Term frequency and inverse document frequency (TF-IDF) offers several benefits in text analysis. Firstly, it considers the importance of words by assigning higher weights to terms that are both frequent within a document and rare in the overall corpus. This allows the model to capture the discriminative power of words and identify key features for analysis. Secondly, Term frequency and inverse document frequency (TF-IDF) reduces the influence of commonly occurring words (such as "the" or "and") that are not informative for distinguishing documents. By focusing on rare and meaningful terms, TF-IDF enhances the ability to extract relevant information. Additionally, Term frequency and inverse document frequency (TF-IDF) can be particularly useful in tasks such as document classification, information retrieval, and text clustering, where it helps improve the accuracy and effectiveness of these analyses. Overall, applying Term frequency and inverse document frequency (TF-IDF) in text analysis provides a valuable approach for capturing the significance of words and improving the quality of results.

B. Model Fitting

This study utilizes the random forest and support vector machine algorithms to build a text classification model. The model is trained using the training set and evaluated using cross-validation scores. The performance of the model on the training set is assessed, and the test set is used to calculate and evaluate accuracy, recall, precision, and F1-score. The results of these evaluations are presented in Tables II and III, respectively.

C. Support Vector Machine (SVM)

SVM is a powerful technique for sentiment analysis. It finds an optimal hyper plane to separate classes, handles high-dimensional feature spaces, and captures complex relationships between words and sentiments. SVM is well-suited for tasks where feature space and non-linear relationships play a crucial role. Support Vector Machines (SVM) is a highly versatile and powerful machine learning algorithm that is widely used for both classification and regression tasks. It has proven to be particularly effective in handling complex and high-dimensional data, making it a popular choice in various applications such as text classification, image recognition, and bioinformatics. One of the key strengths of SVM lies in its ability to handle non-linear decision boundaries through the use of kernel functions. By employing these kernels, SVM can capture intricate relationships and patterns in the data. The fundamental principle behind SVM is to find the optimal hyperplane that maximizes the separation between different classes in the feature space. This hyperplane, also known as the decision boundary, aims to achieve the largest margin between the classes, providing a robust and well-generalized solution. The data points that lie closest to the decision boundary, known as support vectors, play a crucial role in defining the decision boundary and overall classification performance. SVM offers several advantages, including its ability to handle noisy datasets and its robustness against overfitting. By focusing on maximizing the margin while minimizing the classification error, SVM can provide good generalization to unseen data. Additionally, SVM allows for the incorporation of different kernel functions, such as linear, polynomial, Gaussian (RBF), or sigmoid, offering flexibility to adapt the algorithm to various data types and problem domains. To train an SVM model, a convex optimization problem is solved, which can be efficiently tackled using various optimization algorithms. The result is a well-optimized model that can accurately classify new instances based on learned patterns and relationships in the training data. Overall, SVM is a versatile and reliable machine learning algorithm that has gained popularity due to its strong performance and ability to handle complex data. Its effectiveness in capturing non-linear relationships and generalizing well to new data makes it a valuable tool for data scientists and researchers in various fields.

D. Random Forest

Random Forest is another effective approach for sentiment analysis. It combines multiple decision trees to make predictions, offering robustness against over fitting and handling imbalanced class distributions. Random Forest is suitable for tasks where interpretability, ensemble learning, and handling complex datasets are important considerations. Random Forest is a highly effective and popular machine learning algorithm that is widely used for both classification and regression tasks. It is known for its ability to handle complex datasets with high-dimensional features and provide accurate predictions. Random Forest gets its name from the ensemble of decision trees it employs to make predictions. The algorithm works by creating a multitude of decision trees, each trained on a random subset of the original dataset. These individual decision trees are constructed by randomly selecting features at each node and making splits based on the best possible criteria, such as Gini impurity or information gain. By combining the predictions of multiple decision trees, Random Forest reduces the risk of overfitting and improves the overall accuracy and robustness of the model. One of the key advantages of Random Forest is its ability to handle both categorical and continuous features without requiring extensive preprocessing. It can handle missing data and maintain good predictive performance even in the presence of noisy or irrelevant variables. Random Forest also provides valuable insights into feature importance, allowing users to identify the most influential features in the prediction process. This information can aid in feature selection and understanding the underlying relationships within the dataset. The algorithm is computationally efficient, as the training of individual decision trees can be performed in parallel. It can handle large datasets with numerous features, making it suitable for both small and big data applications. Random Forest is a versatile algorithm that can be applied to various domains, including finance, healthcare, marketing, and more. Its robustness, accuracy, and ability to handle complex data make it a popular choice among data scientists and researchers.

VI. RESULTS AND DISCUSSIONS

Result shows the training set and testing set and their respective sum. Table 1 shows that dataset text segmentation tool and data stop words removal techniques have been applied to remove all the unnecessary information from the text dataset. The data set was collected manually. In summary, Random Forest is a powerful machine learning algorithm that leverages an ensemble of decision trees to provide accurate predictions. Its ability to handle complex data, handle missing values, and provide insights into feature importance makes it a valuable tool for a wide range of applications.

Table 1: Dataset

Dataset	Positive Data	Negative Data	Sum
Training Set	4804	4794	9598
Testing Set	1194	1205	2399
Sum	5998	5999	11997

Table 2 represents the testing results of random forest and support vector machine for five iterations.

Table 2: Testing Results for Five Iterations

Algorithm	First Score	Second Score	Third Score	Fourth Score	Fifth Score	Average
Random Forest	0.78518	0.78565	0.78596	0.79018	0.78112	0.78561
SVM	0.80227	0.79971	0.80190	0.80680	0.80044	0.80222

Table 3 demonstrates the accuracy, recall, and precision, F1-score, random forest, and support vector machine.

Table 3: Accuracy, Recall, Precision, F1-score for Random Forest and SVM

Algorithm	Accuracy	Recall	Precision	F1-score
Random Forest	0.78564	0.78564	0.78737	0.78527
SVM	0.80394	0.80394	0.80654	0.80347

E. Comparative Analysis

Based on the experimental findings presented above, it is evident that the model constructed using the support vector machine (SVM) algorithm outperforms the random forest algorithm in terms of both fitting degree and generalization ability. These results have led the authors of this paper to select the SVM algorithm as the preferred choice for building the model. By considering the performance metrics, such as accuracy, recall, precision, and F1-score, it is apparent that the SVM algorithm demonstrates superior capabilities in accurately capturing the patterns and characteristics of the sentiment analysis task at hand. The decision to adopt SVM is based on its ability to effectively handle complex feature spaces, capture intricate relationships between words and sentiments, and achieve a high degree of classification accuracy on both the training set and the test set. Through this selection, the paper aims to leverage the strengths of the SVM algorithm and harness its potential to improve sentiment analysis outcomes. Random Forest is a powerful ensemble learning algorithm that combines multiple decision trees to make predictions. It demonstrated a solid performance with an accuracy of 0.78564. Random Forest's ability to handle complex datasets, handle missing values, and provide insights into feature importance makes it a valuable choice for various applications. On the other hand, SVM (Support Vector Machines) proved to be even more accurate with a score of 0.80394. SVM is a versatile algorithm known for its effectiveness in handling high-dimensional data and capturing complex relationships. It finds the optimal hyper plane to separate different classes in the feature space, aiming to maximize the margin while minimizing classification errors. This allows SVM to generalize well to new and unseen data.

V. CONCLUSION

This comprehensive research paper has provided an overview of sentiment analysis, covering various machine learning techniques used in the field. It discussed preprocessing techniques, feature extraction methods, and explored six machine learning techniques: Naive Bayes, Support Vector Machines, Decision Trees, and Random Forests, Neural Networks, and Unsupervised Learning approaches. The paper also delved into model training, evaluation metrics, and highlighted the importance of comparative analysis. By understanding these concepts and techniques, researchers and practitioners can effectively apply sentiment analysis in different domains and contribute to advancements in sentiment analysis algorithms and methodologies. Random Forest is suitable for tasks where interpretability, ensemble learning, and handling complex datasets are important considerations. Upon conducting an in-depth analysis of the model, it has been observed that a significant portion of the comment data exhibits a predominant negative sentiment inclination. Specifically focusing on the negative data subset, it becomes evident that a considerable proportion of the comments express sentiments that lean towards negativity. This finding sheds light on the prevailing sentiment pattern within the analyzed dataset, highlighting the prominence of negative sentiments among the comments. By recognizing this notable tendency towards negativity, further exploration and examination of the underlying factors and themes contributing to this sentiment can be pursued. The identification of such trends in sentiment distribution serves as valuable insight, enabling researchers and analysts to gain a comprehensive understanding of the sentiment landscape within the comment dataset and potentially uncover relevant patterns and themes driving the prevailing negative sentiments.

ACKNOWLEDGEMENT

The authors would like to thank the two anonymous reviewers who have provided valuable suggestions to improve the article.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends® in information retrieval*, vol. *2*, no. 1–2, pp 1-135, 2008.

[2] B. Liu, "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp 1-167, 2012.

[3] E. Cambria and A. Hussain, "Sentic Computing: Techniques, Tools, and Applications", *Springer*, 2012.

[4] S. Kiritchenko and S. Mohammad, "Examining the use of sarcasm on Twitter for sentiment analysis", *Proceedings of the conference on empirical methods in natural language processing*, 2, 2018, pp.7-12.

[5] C.D. Manning and H. Schütze, "Foundations of statistical natural language processing", *MIT Press*, 1999.

[6] S. Bird, E. Klein and E. Loper, "Natural language processing with Python", *O'Reilly Media Inc*, 2009.

[7] M. F. Porter, "An algorithm for suffix stripping", *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[8] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey", *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014.

[9] S. Bird, E. Klein and E. Loper, "Natural language processing with Python", *O'Reilly Media Inc*, 2009.

[10] B. Pang and L. Lee, "Opinion mining and sentiment analysis", *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[11] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", *In European conference on machine learning*, *Springer*, 1998, pp. 137-142.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information processing & management*, vol. 24, no. 5, pp. 513-523, 1988.

[13] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *In Advances in neural information processing systems*, pp. 3111-3119, 2013.

[14] R.W.Picard, "Affective computing", MIT press, 2000.

[15] A. Alfarrarjeh, S. Agrawal, S. H. Kim and C. Shahabi, "Geo-Spatial Multimedia Sentiment Analysis in Disasters", *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, pp. 193-202. doi: 10.1109/DSAA.2017.77.

[16] M. Maia, A. Freitas and S. Handschuh, "FinSSLx: A Sentiment Analysis Model for the Financial Domain Using Text Simplification", *IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018, pp. 318-319. doi: 10.1109/ICSC.2018.00065.

[17] S. Vanaja and M. Belwal, "Aspect-Level Sentiment Analysis on ECommerce Data", *International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018, pp. 1275-1279. doi: 10.1109/ICIRCA.2018.8597286.

[18] A. J. Nair, G. Veena and A. Vinayak, "Comparative study of Twitter Sentiment on COVID - 19 Tweets", *5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1773-1778. doi: 10.1109/ICCMC51019.2021.9418320.

[19] C. Whissell, "Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language", *Psychological Reports*, vol. 105, no. 2, pp. 509–521, 2009. doi:10.2466/pr0.105.2.509- 521.

[20] S. Park, and Y. Kim, "Building thesaurus lexicon using dictionary-based approach for sentiment classification", *IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2016. doi:10.1109/sera.2016.7516126.

[21] B. Pang, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", *Proc. EMNLP*, Philadelphia. PA, USA, July 2002.

[22] S. Hemalatha and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning", *International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp.700-704. doi: 10.1109/ICCS45141.2019.9065812.

[23] Y. Sun, X. G. Zhou and W. Fu, "Unsupervised Topic and Sentiment Unification Model for Sentiment Analysis", *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 49, no. 1, pp. 102-108, 2013.

[24] Z. G. Jin and Y. Yang, "A semi-supervised short text sentiment analysis model based on social relationship strength", *Journal of Harbin Institute of Technology*, vol. 51, no. 5, pp. 50-56, 2019.

[25] T. T. Nguyen and A. G. Kravets, "Analysis of the social network facebook comments", 7th International Conference on Information, Intelligence, Systems & Applications (IISA), 2016, pp. 1- 5. doi: 10.1109/IISA.2016.7785412.

[26] A. A. Gamova, A. A. Horoshiy and V. G. Ivanenko, "Detection of Fake and Provokative Comments in Social Network Using Machine Learning", *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2020, pp.309-311. doi: 10.1109/EIConRus49466.2020.9039423.

[27] S. Mestry, H. Singh, R. Chauhan, V. Bisht and K. Tiwari, "Automation in Social Networking Comments With the Help of Robust fastText and CNN", *1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 2019, pp.1-4. doi: 10.1109/ICIICT1.2019.8741503.

[28] R. Meena and V. T. Bai, "Study on Machine learning based Social Media and Sentiment analysis for medical data applications", *Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019, pp.603-607. doi: 10.1109/ISMAC47947.2019.9032580.

[29] T. A. Khan *et al*., "An Implementation of Electroencephalogram Signals Acquisition to Control Manipulator through Brain Computer Interface", *IEEE International Conference on Innovative Research and Development (ICIRD)*, Jakarta, Indonesia, 2019, pp.1-6. doi: 10.1109/ICIRD47319.2019.9074722.

[30] T. A. Khan, M. Alam, Z. Shahid and M. M. Suud, "Prior investigation for flash floods and hurricanes, concise capsulization of hydrological technologies and instrumentation: A survey", *2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS)*, Bangkok, Thailand, 2017, pp. 1-6. doi: 10.1109/ICETSS.2017.8324170.

[31] T. A. Khan, M. Alam, K. A. Kadir, Z. Shahid and M. Mazliham, "Artificial Intelligence based prediction of seizures for Epileptic Patients: IoT based Cost effective Solution", *2019 7th International Conference on Information and Communication Technology (ICoICT)*, Kuala Lumpur, Malaysia, 2019, pp.1-5. doi: 10.1109/ICoICT.2019.8835350.

[32] T. A. Khan, M. M. Alam, Z. Shahid and M. M. Su'ud, "Prior Recognition of Flash Floods: Concrete Optimal Neural Network Configuration Analysis for Multi-Resolution Sensing", *in IEEE Access*, vol. 8, pp. 210006-210022, 2020. doi: 10.1109/ACCESS.2020.3038812.

[33] T.A. Khan, M. Alam, K. Kadir, Z. Shahid, and M.S. Mazliham, "Prior Determination of Flash Floods: Artificial Intelligence Based Predictive Analysis using Modified Cuckoo Search", *J. Comput. Theor. Nanosci*, vol. 17, no. 2-3, pp. 990–995, Feb 2020.

[34] T.A. Khan, M. Alam, Z. Shahi, and M.S. Mazliham, "Performance comparison of SVM and its variants for the early prognosis of breast cancer", Sukkur IBA Journal of Computing and Mathematical Sciences, [S.l.], vol. 3, no. 2, pp. 1-8, mar. 2020. ISSN 2522-3003.

[35] T.A. Khan, S. Ahmed, S.S.A. Rizvi, S. Ahmad and N. Khan, "Electromyography Based Gesture Recognition: An Implementation of Hand Gesture Analysis Using Sensors", *Sir Syed University Research Journal of Engineering and Technology,* vol. 12, no. 1, pp. 70-77, July 2022. https://doi.org/10.33317/ssurj.424.

[36] H. Shelar, G. Kaur, N. Heda and P. Agrawal, "Named Entity Recognition Approaches and Their Comparison for Custom NER Model", *Science & Technology Libraries*, vol. 39, no. 3, pp.324–337, 2020. doi:10.1080/0194262X.2020.1759479.

[37] H. Panoutsopoulos, C. Brewster and B. Espejo-Garcia, "Developing a model for the automated identification and extraction of agricultural terms from unstructured text", *IOCAG* 2022. doi:10.3390/iocag2022-12264.

[38] Y. Huo, Y. Su and M. Lyu, "LogVm: Variable Semantics Miner for Log Messages", *IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, Charlotte, NC, USA, 2022, pp.124-125. doi: 10.1109/ISSREW55968.2022.00053.

[39] J. Wan, M. Liu, J. Yi and X. Zhang, "Detecting spam webpages through topic and semantics analysis", *2015 Global Summit on Computer & Information Technology (GSCIT)*, Sousse, Tunisia, 2015, pp. 1-7. doi: 10.1109/GSCIT.2015.7353328.

[40] Y. Lim, K.W. Ng, P. Naveen, and S.C. Haw, "Emotion Recognition by Facial Expression and Voice: Review and Analysis", *Journal of Informatics and Web Engineering*, vol. 1, no. 2, pp .45-54, 2022.

[41] C.Y. Seek, S.Y. Ooi, Y.H. Pang, S.L. Lew and X.Y. Heng, "Elderly and Smartphone Apps: Case Study with Lightweight MySejahtera", *Journal of Informatics and Web Engineering*, vol. 2, no. 1, pp .13-24, 2023.