Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study

Wesam Ahmed
Information Technology Department
Faculty of Computers and Information
South Valley University
Hurghada, Egypt
wesam.elbaz@fcih.svu.edu.eg

Noura A.semary
Information Technology Department
Faculty of Computers and Information
Menoufia University
Menoufia ,Egypt
noura.semary@ci.menofia.edu.eg

Khalid M. Amin
Information Technology Department
Faculty of Computers and Information
Menoufia University
Menoufia, Egypt
k.amin@ci.menofia.edu.eg

Mohamed Hammad
Information Technology Department
Faculty of Computers and Information
Menoufia University
Menoufia, Egypt
mohammed.adel@ci.menofia.edu.eg

Abstract:* The term "machine learning" refers to a sort of artificial intelligence (AI) that empowers software applications to enhance their predictive capabilities without explicit programming for such purposes. In order for machine learning algorithms to anticipate future output values, they require past data as input. In terms of scope, this research falls under sentiment analysis. The latter field is becoming increasingly active in terms of extracting people's opinions on issues related to politics, economics, and social issues. The purpose of sentiment classification is to categorize users' opinions as neutral, positive, or negative based on textual input alone. Despite these advantages, the accuracy and effectiveness of sentiment analysis are compromised by the obstacles encountered in the field of natural language processing (NLP). Recent research has shown that machine learning algorithms can assist in NLP. In this research, we investigate a range of machine-learning strategies for solving sentiment analysis challenges. Two datasets were analyzed with the models based on the term frequency-inverse document frequency (TF-IDF). A comparison study was conducted between each of the models to determine how they performed in experiments. Regarding accuracy and F1 score, logistic regression performs better than other algorithms.

Keywords: *TF-IDF; AI; sentiment analysis; machine learning; NLP.*

I. Introduction

Social media platforms provide valuable data for sentiment analysis (SA). The constant expansion of social networks has resulted in far more intricate and interconnected data. It is now more important than ever for businesses to pay close attention to the voice of the customer (VoC) in order to enhance their products. Product managers require insights to help them develop a solid product roadmap; it is about giving customers what they actually want rather than what businesses believe they need [١] .

Using sentiment analysis, which is often called "opinion mining" in the field of natural language processing (NLP), you can find the sentiment scores of textual data.It examines people's feelings and attitudes toward products, events, digital content, government agencies, and issues [2]. While huge amounts of opinion data can give in-depth insights into overall sentiment, processing them takes a long time. The process of reviewing huge volumes of material may be both laborious and challenging. Additionally, certain texts may possess considerable length and complexity, presenting various lines of reasoning for distinct sentiments, thereby rendering the overall sentiment difficult to comprehend readily [3]. This paper will solve the previous problems by using machine learning models and feature extraction techniques.

The following are the key contributions to the proposed work:

- The paper provides insights into which techniques work best for Twitter sentiment analysis and uses the Twitter dataset because this dataset can help identify patterns and trends that may not be apparent through other data sources.
- Comparing four ML techniques and the strengths and weaknesses of each technique using TF-IDF feature extraction, which is simple, efficient to compute, and focuses on the more informative words in the document.
- It has been demonstrated that using logistic regression is a very important technique in Twitter sentiment classification because of its robustness and interpretability, which means the ability to handle a large number of input features without overfitting and provide insights into the importance of each feature in the prediction process.

The remaining sections of this work are organized as follows: The second section provides context for the topic of this research. In the third section, pertinent literature is discussed. The experimental results are examined in the fourth section, followed by the conclusion.

II. Background

A. Sentiment Analysis

The phrase "sentiment analysis" was first introduced in 2001 through research endeavors focused on forecasting and uncovering market sentiment through the examination of evaluative texts [4]. Many application domains, including biomedicine, government, and business, have found sentiment analysis highly useful. This term refers to the task of categorizing reviews into positive and negative sentiment polarities. In terms of granularity, sentiment mining or opinion mining can be performed at several levels, including feature, sentence, and document levels. At the document level, one seeks to learn the author's general opinion, for instance, in a movie or a product review. The sentiment or polarity of a single sentence can be determined by word-level analysis. The documents and statements in question do not explicitly identify the subject to which the author is directing their sentiment, whereas sentences can address various topics and entities. Due to these factors, fine-grained analyses are required. It is primarily possible to divide sentiment analysis methods into three types:

- Lexicon-based approaches
- Machine learning-based approaches
- Hybrid approaches

Lexicon-based sentiment analysis techniques [6]: The primary methodologies are dictionary-based and corpus-based. The valence dictionary assigns positive or negative (and sometimes neutral) labels to words in texts. By counting the number of positive and negative words in the text and combining their values mathematically, we can derive an overall sentiment score.

Machine learning techniques [7] use a dataset whose sentiment is already known to determine sentiments for newly unlabeled data. The training of these algorithms can be accomplished by the utilization of many methodologies, including neural networks, decision trees, and logistic regression.

A hybrid method [8] is a method that exploits both knowledge and statistical-based methods to detect polarity. As a result of the lexicon-based approach and machine learning approach, high accuracy, and stability are inherited.

B. Machine Learning

ML is a powerful tool for sentiment analysis in various applications and domains, which makes it an increasingly popular choice for businesses and organizations looking to gain insights from text data [9]. Here are some examples of applications and domains where ML is used in Twitter sentiment analysis: 1- Social research: to analyze social issues like racism and sexism. 2- Customer service: to identify customer sentiment and provide personalized responses. 3- Political analysis: to analyze tweets related to political events like elections and debates. ML techniques can be trained on specific domains or industries, which allows for more accurate sentiment analysis tailored to a particular context, and this is one of the reasons for using ML over other approaches [10][11]. The following steps are used to classify the input textual sentiment into positive, Negative, and neutral sentiment. The input text is cleaned and pre-processed, then the input text is converted into a set of features that can be used to train the machine-learning model and a ML algorithm is trained on a labeled dataset of examples with known sentiment polarity. To evaluate the model's accuracy and performance, a separate dataset is used.

C. Linear Classifiers

1) Naive Bayes: Its simplicity and reasonably good performance make Naive Bayes (NB) a popular classifier [10]. The assumption of NB learning is that attributes or features are independent. Real-world problems do not always follow this rule, so sometimes it is not preferred even though it can produce an accurate, reliable result. It is a simple and fast classifier that can work well with high-dimensional datasets.

Here is the formula that can be used to denote Bayes classifiers [10]:

$$\mathrm{P}\,(\mathrm{C} \mid \mathrm{X}) = \frac{p(X\mid C)\,p(C)}{p(X)} \qquad (1)$$

where:

- P(C | X) describes the posterior probability of class C given input features X.
- p(X | C) represents the conditional probability of observing input features X given class C.
- p(C) represents the prior probability of class C.
- p(X) describes the probability of observing input features X.

2) Logistic Regression: A logistic regression model estimates the likelihood of a parameter having a certain value based on the values of the other parameters. We can use this technique when the interpretability of the model is important but it may not perform well when the relationship between the independent and dependent variables is nonlinear. There are two possible values for a parameter (binary) or many more (categorical). It is not inherently a classifier, but it can be used as one if threshold values are chosen that determine if a parameter belongs to one or another class. There are many ways to choose the threshold, such as domain knowledge, and another way is to use the common default threshold value of is 0.5 [12].

3) Support Vector Machine (SVM): is a contemporary supervised machine learning approach, which is highly accurate but computationally expensive. SVM learns from a set of data samples classified into one of two categories, then attempts to assign new data to the appropriate category. In n-dimensional vector space, the technique endeavors to locate a hyperplane that effectively partitions the data instances into two distinct sides, while maximizing the margin between them. Thus, (n-1) dimensions will be used to define this hyperplane. [13].

4) Random Forest: In random forests (RFs), a multitude of decision trees are constructed at training time during the learning process for tasks such as classification and regression. They can handle both categorical and continuous variables but are slow to train on large datasets. The output of a random forest is a class selected by most trees. Although they frequently perform better than random forests, gradient-boosted trees are more accurate. But data features can impact how well they work [14].

D. TF-IDF Feature Extraction of Textual Data

The utilization of phrase Frequency-Inverse Document Frequency (TF-IDF) is a prevalent method employed to assess the significance of a phrase within a document and its relative rarity across the entire corpus. In the ML sentiment analysis process resulting TF-IDF scores for each term in each tweet are combined into a feature vector that represents the tweet. This feature vector can be used as input to ML algorithms [14].

E. Evaluation Metrics:

Accuracy and F-score are used for evaluating the performance of machine learning models. Accuracy: Calculated by dividing all the correct predictions by the total predictions. F-score: It is the harmonic mean of precision and recall [14].

III. RELATED WORK

In this study, different methods of sentiment analysis are reviewed that can be used in future empirical research, and Table I shows a summary of related work.

Madanjit et al. [17] provided an examination of the behavior of agricultural players in the Indian region during the epidemic. With the Linear Discriminant Analysis (LDA) algorithm. The findings indicate that the COVID-19 pandemic had a significant impact on the socioeconomic status of agricultural workers. This study's weakness is that the majority of farmers may not have social media profiles.

Qiu *et al.* [18] investigated sentiment analysis and machine learning-based short-term stock trend prediction. When applied to models such as RF, SVM, Gradient Boosting Decision Trees (GBDT), K-Nearest Neighbors (KNNs), Decision Trees (DT), NB, and Logistic Regression (LR), the adjusted sentiment index can improve the capacity to predict stock trend direction. According to the adjusted sentiment index and market indicators, seven of the eight models had the highest F1-score and precision.

Atteveldt *et al.* [19] a validation set of Dutch economic headlines was used to evaluate machine learning, hand annotation, multiple dictionaries, both conventional and deep learning methods, and crowd coding. Experimental results show that trained humans or crowds achieve the best performance.

Sudhanshu *et al.* [20] analyzed sentiments based on age groups through dictionary-based techniques and machine learning, and the impact of gender and age on user reviews was investigated. The accuracy of over-50 age group has the best accuracy as compared to all other age groups.

In their study, Sajeetha *et al.* [21] conducted sentiment analysis (SA) to investigate five different methodologies. These methodologies include the supervised machine learning-based approach, the K-means with Bag of Word (Bow) approach, the K-modes with Bow approach, the hybrid approach, and the lexicon-based technique. Results have shown a significant improvement in accuracy when using pre-processed corpus and nouns as corpus. Among all feature representation techniques tested, fast text stands out as the most efficient.

TABLE I : SUMMARY OF RELATED WORK

Year	Study	Research Work	Dataset	Advantages	Limitations
2022	M.Singh et al [12]	Twitter sentiment analysis	Twitter	Real-time insights	Twitter users may not be representative of all tourists in Thailand
2022	Y. Qiu et al [13]	Sentiment analysis-based forecasting of short-term stock trends	market indicators and the related reviews on Eastmoney.com.	The accuracy is high value	Limited timeframe and can focus in long-term stock also
2021	W.van Atteveldt et al [14]	Crowd-Coding, Dictionary Approaches, ML Algorithms, and Manual Annotation are all compared.	NEWS	work is interdisciplinary	Limited scope
2020	Sudhanshu et al [15]	find different insights on how age and gender affect SA	reviews on books from Facebook users	uses a large dataset of tweets	Limited contextual information
2019	Sajeetha et al [16]	Sentiment Analysis in Tamil Texts	Tamil Texts	evaluates different feature representation techniques	Limited sample size

IV. SIMULATION RESULTS

The purpose of this section is to introduce different models related to datasets. The sentiment polarity analysis we conducted relied on two datasets. In the first dataset, tweets were labelled as having either positive or negative sentiment, while the second dataset contained tweets labelled positively or negatively. Inputs to the classification algorithms are prepared using the TF-IDF approach. The scikit-learn library provided the vectorizer class for TF-IDF. In this experiment, we trained and evaluated four models, each of which had been preprocessed with TF-IDF before training (LR, NB, RF, and SVM). These strategies were compared to advance the state of the art in sentiment analysis.

A. Datasets

Different datasets were gathered from different sources. Machine learning models have been compared comprehensively in sentiment analysis based on the results. Below is a description of these datasets:

- We obtained the Twitter US Airline dataset from [22], and Table II shows the first three examples.
- Twitter US Airlines was obtained from [17]. The tweets were labeled (0 = negative, 4 = positive). The choice of 4 as the positive class label value was made to indicate that is strongly positive and 4 is a higher number than 0, so 0 is negative.
- Generic airline tweets are text files that are scraped from Twitter by using Tweepy [23]. The polarity of tweets had already been labeled (0 = negative, 4 = positive).

In Figure 3 and Figure 4, the 100 most frequent words in the two datasets are described in subsection datasets. This figure shows how easy it is to identify topics. On the top, the most frequently used words in US airline tweets are displayed, and on the bottom, they are displayed in generic tweets.

The experiment was conducted using the "text" and "sentiment" fields.

TABLE II: Examples of the raw US airline dataset.[22]

Id	Sentiment	Reason	User	count	Text
5.70E+17	Positive		jnardino	0	@VirginAmerica plus you've added commercials to the experience... tacky.
5.70E+17	negative	Bad - Flight			@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &amp; they have little recourse
5.70E+17	negative	Can't - Tell			@VirginAmerica and it's a really the big bad thing

B. Sentiment Classification

In the preprocessing stage, data was cleaned, and features were extracted. At the training stage, machine learning models were used. This study is primarily focused on evaluating machine learning models. With TF-IDF, all of them were tested.

The cleaning step is the first preprocessing step that removes special characters. By using regular expressions, we can effectively remove special characters from the input text, ensuring that the subsequent analysis and classification processes focus on the essential content and features of the text, stop words, and URLs from the tweet message and convert the text to lowercase. Figure 2 shows the steps of this process. Both regular expressions and the Beautiful Soup module are used to clean the text and reduce the effectiveness of sentiment analysis. After this step, convert sentences to single words and return them to their base form by using lemmatization. The TF-IDF was used at this point to convert sentences into feature vectors. The ML algorithms used these feature vectors as inputs. Then class label values {4, 0} were used as (4 = positive, 0 = negative).

Input Raw Tweet
Pre-Processing Algorithm
Output Processed Tweets

Fig. 1: Data pre-processing stage.

In sentiment analysis, the dataset is split 60:40 between train and test sets. This step is dependent on the approach used to classify sentiments. During this step, machine learning classifiers were trained to recognize whether each tweet contained a positive or negative tone. The classifier algorithms learn from labelled data, in our case, text tweets, during the training stage. The classifier must, therefore, be able to classify new, unlabeled tweets during the testing stage. By using the accuracy and F1-score, each classifier was then evaluated.

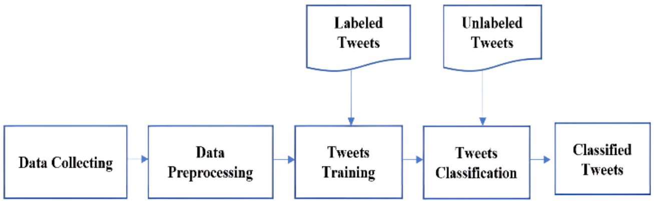

Fig. 2. The proposed workflow of Twitter sentiment classification.

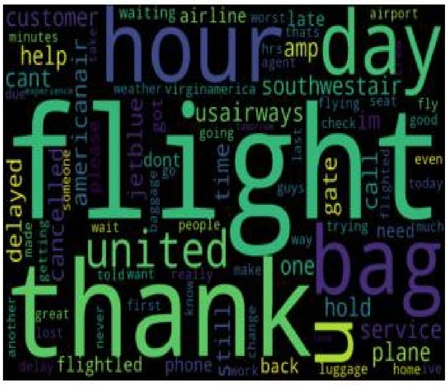

Fig. 3. A word cloud depicting the subjects of the Twitter dataset.

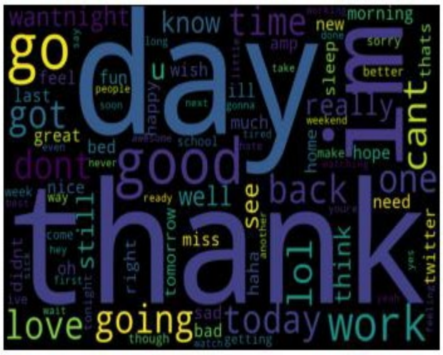

Fig. 4. A word cloud depicting the subjects of the generic dataset.

C. Results

To conduct the tests, we used the Sklearn library [24]. Using the datasets, experiments were conducted using the models (LR, SVM, NB, and RF), and TF-IDF feature extraction and Count Vectorizer performed the tasks of tokenizing and counting.

In the model implementation, hyperparameter tuning is used because the value of parameter C represents the inverse of regularization strength. By choosing the correct C parameter, it prevents the logistic regression from overfitting the training data. As many features are used in this model, it is already unlikely to be underfit. Grid Search cross-validation is used to test the model with various C parameters ranging from (0.001 to 100). The range of C (inverse of regularization strength) from (0.001 to 100) is a common range used in machine learning algorithms, particularly in classification problems. The value that gives the best performance on the CV will be used. Some classifiers are trained (LR, Random Forest , SVM, and Naive Bayes). They will be validated with the generic tweet test set and the complete US airline Twitter dataset. The accuracy and weighted average of the F1 score are used to analyze the models' performance.

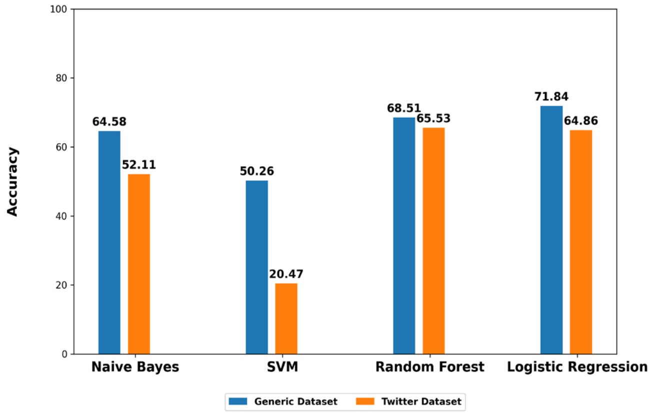

Fig. 5. Comparison of accuracy by different classifiers on the generic tweet and US airline tweet.

According to Figure 5, the accuracy percentage of the logistic regression algorithm is 71.84% on generic tweets and it performs the best. Followed by a random forest classifier with an accuracy of 68.51% on the generic tweet dataset and 65.53% on the US airline dataset. In contrast, SVM performs the worst. The accuracy is 50.26% on the generic tweet dataset, while it is only 20.47% on the US airline dataset. SVM performs the worst with an accuracy of 50.26% on the generic tweet dataset and only 20.47% on the US airline dataset. This suggests that SVM may not be an effective choice for sentiment classification on these datasets.

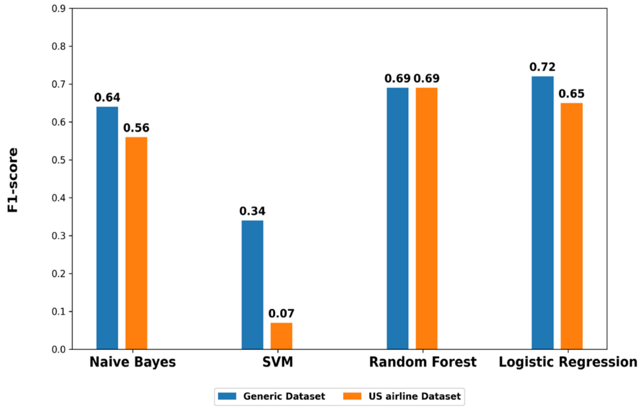

Fig. 6. Comparison of F1 scores by different classifiers on the generic tweet and US airline tweet.

From Figure 6, the weighted average of the F1 score value has the same accuracy. Logistic regression performs the best, with 0.72 on the generic tweet test set and 0.68 on the US airline dataset. It is followed by a random forest classifier with 0.69 on the generic tweet and 0.69 on the US airline dataset. The Naive Bayes algorithm performs 0.64 on the generic tweet dataset. This suggests that the Naïve Bayes model may not be able to capture the more complex language patterns and nuances of sentiment in the dataset. SVM shows the worst F1score value of 0.34 on generic tweet dataset and 0.07 on US airline dataset. This suggests that logistic regression is effective at capturing the complex language patterns and nuances of sentiment in these datasets. The random forest also performs well on both datasets, indicating that it is a viable alternative to logistic regression.

Overall, the results presented in Figures 5 and 6 suggest that logistic regression and random forest classifiers are effective choices for sentiment classification, especially in the generic tweet dataset because the size of dataset is larger than on the US airline dataset, while Naive Bayes and SVM may perform less optimally.

TABLE III: COMPARISON OF ACCURACY BETWEEN PROPOSED MODEL AND PREVIOUS WORK MODELS

Paper	LR	SVM
Proposed Model	72.%	50.26%
Y. Qiu[18]	63.27%	63.26%
S. Kumar[20]	-	71%
S. Thavareesan[21]	71%	68%

From Table III, the LR algorithm of the proposed model outperforms other work.

V. CONCLUSION

This study describes the fundamental machine learning models and related approaches applied to Twitter data sentiment analysis. The insight, along with the outcomes of our experiments, provides us with a thorough understanding of the use of machine learning models for sentiment analysis. As a result of the experiments, logistic regression outperformed other models when it came to sentiment analysis, with a score of 71.84% accuracy on a generic tweet, and a score of 64.86% on a US airline dataset. The SVM performs the worst in both accuracy and weighted average of the F1 score. On generic tweets, SVM has an accuracy of 50.26%, but only 20.47% on US airline tweets. In our future work, we will concentrate on deploying hybrid methodologies, such as deep learning models, to improve sentiment classification accuracy.

REFERENCES

[1] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, Springer,pp. 1–168, 2012.

[2] C. C. Aggarwal and C. Zhai, "A Survey of Text Classification Algorithms.Boston", MA: Springer US, pp. 163–222, 2012.

[3] B. Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions", 2nd ed. Cambridge University Press, pp.1- 452 , 2020.

[4] M Wankhade, A.C.S Rao and , C. A Kulkarni," survey on sentiment analysis methods, applications, and challenges," Artif Intell Rev 55, pp.5731–5780, 2022.

[5] T. N Fatyanosa, , F. A. Bachtiar,"Classification method comparison on Indonesian social media sentiment analysis," International Conference on Sustainable Information Engineering and Technology (SIET), pp.310-315, 2017.

[6] B .Bhavitha., A.P .Rodrigues, , N.N .Chiplunkar," Comparative study of machine learning techniques in sentimental analysis", In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10–11, pp. 216–221, 2017.

[7] P.K.Mallick,V.E.Balas,S.Mishra,andM.K.Mishra "Sentiment Analysis and Evaluationof Movie Reviews Using Classifier",Advances in Intelligent Systems and Computing,Cognitive Informatics and Soft Computing Proceeding of CISC , Vol 768 ,pp.53-59,2017.

[8] W.Medhat, A.Hassan, H. Korashy, "Sentiment analysis algorithms and applications: A survey",Ain Shams Eng. J., pp.1093–1113, 2014.

[9] S. Zad, M. Heidari, J. H .Jones, O. Uzuner," A Survey on Concept-Level Sentiment Analysis Techniques of Textual Data", 2021 IEEE World AI IoT Congress (AIIoT),pp. 0285-0291, 2021.

[10] B. Liang, H. Su, L. Gui, E. Cambria, R. Xu,"Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks", Knowledge-Based Systems,Vol 235, 2022.

[11] A. P Rodrigues , R. Fernandes , A .Aakash, B .Abhishek, A. Shetty, K .Atul, K. Lakshmanna , and R. Mahammad Shafi," Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques", Hindawi Computational Intelligence and Neuroscience, pp. 1-14, 2022.

[12] M. Costola, O. Hinz, M. Nofer, L. Pelizzon,''Machine learning sentiment analysis, COVID-19 news and stock market reactions'',Research in International Business and Finance, pp.1-14, 2023.

[13] Qi, Y., Shabrina, Z. ,'' Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach''. Social Network Analysis and Mining , 2023.

[14] S. D. Gogula, M. Rahouti ,'' An Emotion-Based Rating System for Books Using Sentiment Analysis and Machine Learning in the Cloud '',applied sciences,pp.1-24, 2023.

[15] U.T. Phan, N. T. Nguyen, D. Hwang,''Aspect-level sentiment analysis: A survey of graph convolutional network methods'',Information Fusion, Pp. 149-172, 2023.

[16] TanL,TanOK,SzeCC,Goh WWB(2023)EmotionalVarianceAnalysis:A new sentiment analysis featureset for Artificial Intelligence and Machine Learning applications.PLoSONE ,pp.1-22, 2023.

[17] M.Singh, A. Singh , P. Singh, and M.Saini "Using Social Media Analytics and Machine Learning Approaches to Analyze the Behavioral Response of AgricultureStakeholdersduringtheCOVID-19 andemic,",MDPI,pp.14-23, 2022.

[18] Y. Qiu , Z. Song and Z. Chen," Short-term stock trends prediction based on sentiment analysis and machine learning,",Soft Computing-springer ,pp.2210-2224, 2022.

[19] W. van Atteveldt, M. A. C. G. van der Velden , M. Boukes "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms", Communication Methods and Measures, Taylor & Francis Group, LLC.,pp121-140,2021.

[20] S. Kumar, M. Gahalawat, P. P. Roy , D. P. Dogra, and Byung-Gyu Kim" Exploring Impact of Age and Gender on Sentiment Analysis Using Machine Learning", mdpi,pp.1-14, 2020.

[21] S. Thavareesan, S. Mahesan," Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation" IEEE 14th International Conference on Industrial and Information Systems (ICIIS), pp.320-326, 2019.

[22] N.C .Dang, M.N. Moreno-García, F.De. La. Prieta. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics* , pp.1-29, 2020.

[23] https://github.com/semaahmed/Generic-dataset.Last Visited 22/5/2023

[24] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. LastVisited 12/12/2022