JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS USING MACHINE LEARNING

[1] Abhishek Gupta, [2] Vandana Bhattacharjee, [3] Aparna U.
[1] Student, [2] Professor, [3] Associate Professor
[1] Department of Electronics & Communication Engineering
[1] MIT Manipal, Udupi, India

Abstract : In the contemporary digital age, sentiment analysis has emerged as a crucial area of research and application owing to the exponential growth of textual data generated across various online platforms such as social media, customer reviews, and news articles. Understanding sentiment from text is pivotal for businesses to gauge customer satisfaction, monitor public opinion, and make data-driven decisions. This research aims to develop a sentiment analysis model using machine learning techniques to classify text data into positive, negative, or neutral sentiments, thereby contributing to the field of natural language processing and information retrieval.

The methodology employed in this research involves several key steps. First, a dataset comprising labeled text samples is collected and preprocessed to remove noise and standardize the text format. Next, a feature extraction technique such as TF-IDF (Term Frequency-Inverse Document Frequency) is applied to represent text data numerically. Subsequently, machine learning algorithms like Support Vector Machines (SVM) and Naive Bayes are employed to train and evaluate the sentiment classification model.

The results of this sentiment analysis demonstrate promising accuracy in sentiment classification across different domains of text data. By leveraging machine learning algorithms and textual feature representations, the developed model achieves robust performance in distinguishing sentiments expressed in text. The significance of this work lies in its practical implications for businesses, market researchers, and social media analysts, enabling them to automate sentiment analysis tasks and derive valuable insights from large volumes of text data efficiently.

In conclusion, this research underscores the effectiveness of machine learning techniques in sentiment analysis, showcasing their potential for real-world applications in sentiment monitoring and opinion mining. The sentiment analysis model developed serves as a valuable tool for extracting sentiment from textual data, aiding decision-making processes in diverse domains. The implementation of this research utilizes Python programming language along with popular libraries such as scikit-learn and TensorFlow for machine learning and natural language processing tasks.

IndexTerms **- Sentiment, tweets, classification, naive bayes, support vector classification, regression**

I. INTRODUCTION

Sentiment analysis has garnered significant interest due to its wide-ranging applications in diverse fields such as marketing, customer feedback analysis, and social media monitoring. The motivation behind undertaking this research stems from several key factors.

Firstly, existing sentiment analysis approaches often encounter challenges in accurately interpreting sentiment from informal and context-dependent text, such as social media posts and product reviews. Many reference papers highlight the limitations of traditional methods in handling sarcasm, irony, and nuanced expressions, leading to inaccuracies in sentiment classification.

Furthermore, the importance of sentiment analysis in today's context cannot be overstated. With the proliferation of online platforms and the rapid generation of user-generated content, organizations require efficient tools to extract actionable insights from large volumes of textual data. Sentiment analysis serves as a crucial component for understanding public opinion, assessing brand sentiment, and improving customer experiences.

The potential result of this research holds substantial significance for both academic research and practical applications. By addressing the shortcomings of existing sentiment analysis techniques, the developed methodology is expected to yield improved accuracy and robustness in sentiment classification across different domains and languages. This enhancement in sentiment analysis capabilities can empower businesses to make data-driven decisions, enhance customer engagement strategies, and gain competitive insights in dynamic market environments.

In summary, the motivation to embark on this sentiment analysis research is driven by the need to overcome the limitations of traditional approaches, harness the power of advanced machine learning techniques, and deliver a scalable sentiment analysis solution with high accuracy and practical utility.

The primary objective of this research is to develop and implement a sentiment analysis system capable of accurately classifying the sentiment expressed in textual data into categories such as positive, negative, or neutral. The main focus is to leverage machine

learning and natural language processing techniques to build a robust model that can automate the process of sentiment classification, thereby facilitating decision-making based on sentiment analysis results.

The secondary objectives of this work include:

- Exploring and comparing different machine learning algorithms and techniques for sentiment analysis to identify the most effective approach.
- Investigating the impact of various text preprocessing techniques (e.g., tokenization, stop-word removal, stemming/lemmatization) on the performance of sentiment analysis models.
- Evaluating the model's performance across different domains and datasets to assess its generalizability and adaptability to various types of textual data.

By achieving these objectives, this research aims to contribute to the advancement of sentiment analysis methodologies and their practical applications, particularly in the context of automated sentiment monitoring and opinion mining from large-scale textual data sources.

The importance of the result of this sentiment analysis research lies in its ability to provide accurate and timely insights into the sentiment conveyed within textual data. By defining target specifications for the sentiment analysis system, we aim to ensure that the developed model meets the desired performance criteria and fulfills the requirements of its intended applications.

The significance of defining target specifications includes:

- Accuracy: Ensuring that the sentiment analysis model achieves a high level of accuracy in classifying sentiments, thereby minimizing misinterpretation of textual data and providing reliable insights.
- Speed and Efficiency: Implementing the sentiment analysis system to process large volumes of textual data efficiently, enabling real-time or near-real-time analysis for timely decision-making.
- Robustness: Designing the model to be robust against noise, variability in language usage, and diverse text sources, thus enhancing its generalizability and applicability across different domains.
- Scalability: Building a scalable sentiment analysis solution capable of handling increasing volumes of data and adapting to evolving business needs and user requirements.
- Interpretability: Ensuring that the sentiment analysis results are interpretable and explainable, allowing stakeholders to understand the basis for sentiment classification decisions and derive actionable insights.

By defining and adhering to these target specifications, the result of the sentiment analysis research will be a reliable and effective tool for extracting sentiment-related information from textual data, ultimately supporting decision-making processes in various domains such as marketing, customer service, and public opinion analysis.

II. LITERATURE REVIEW

This research delves into sentiment analysis methodologies across diverse datasets, focusing on rectifying the prevalent issue of inadequately annotated data for training deep learning models effectively. By systematically contrasting Lexicon-based, Machine Learning, and Deep Learning approaches, with a special focus on sophisticated Transformer models such as BERT and GPT-3, the study offers a comprehensive understanding. Through meticulous evaluation using renowned datasets like IMDB reviews and Sentiment140 for Tweet sentiment analysis, it elucidates the nuanced performance variations among different techniques. Furthermore, the research delves into the influence of textual or tweet features on model efficacy, enhancing our comprehension of sentiment analysis techniques in real-world applications.

A recent work on Twitter movie review sentiment analysis has been done by Kiruthika et al. [14]. They extracted the twitter data using the traditional method of twitter API after building the required application on the developer site. Thereafter, they performed a sentiment analysis of Twitter data about movies using a supervised learning approach.

In their study, Richa Dhanta et al. [1] investigated sentiment analysis on Twitter data, categorizing tweets as positive, negative, or neutral. They employed various machine learning approaches, including logistic regression and Naive Bayesian, after preprocessing the dataset to eliminate noise and irrelevant information. The study evaluated the effectiveness of these methods using metrics like F1 score, accuracy, recall, and precision, revealing promising outcomes.

A research[11] presents three deep learning networks applied to sentiment analysis of IMDB movie reviews, with an equal distribution of positive and negative feedback in the dataset. Among these networks are Convolutional Neural Networks (CNN), Recurrent Neural Networks, and Long Short-Term Memory (LSTM) networks, which are commonly used in natural language processing tasks. The results indicate that CNN achieves superior classification performance in sentiment analysis of movie reviews, demonstrating an accuracy of 88.22%, while RNN and LSTM achieve accuracies of 68.64% and 85.32%, respectively.

Cihan ÇILGIN et al. (2022) [7] conducted a study where they analyzed Twitter data related to COVID-19 by collecting tweets containing various hashtags such as ‘#covid19’, ‘#Covid’, ‘#pandemic’, etc., from January 1 to July 1, 2020, amounting to a total of 60,243,040 tweets. Utilizing VADER, they categorized the emotional tone of these tweets and grouped them into five categories based on their overall sentiment scores. Additionally, they employed Word Clouds to visually represent the most frequently occurring text data each month and used N-grams to gain insights into the content and meaning of the tweets. Notably, while negative tweets dominated in the early stages of the epidemic, there was a shift towards more positive tweets as time progressed.

In their 2021 study, S. Jacob et al. [2] utilized a machine learning-driven clustering method to analyze extensive tweet data, consisting of a qualifying and test collection containing over one lakh results. They aimed to discern the sentiment of each tweet, determining whether it leaned towards positivity or negativity.

A study by Al-Natour and Turetken (2020) [8], which utilized rule-based methods, found that contextual factors like product type and review length impact the accuracy of techniques in detecting genuine sentiment in reviews. Among the various lexicon-based sentiment analysis techniques evaluated, Vader demonstrated slightly better performance.

Wankhade et al. [3] conducted a comprehensive review of sentiment analysis techniques, assessing and comparing various approaches such as Lexicon-based, Machine Learning, Hybrid, and Transfer Learning methods. Their study achieved peak accuracies of 0.90 for IMDB reviews and 0.82 for Sentiment140 utilizing an Attention-based Bidirectional CNN-RNN Deep Model.

In a separate study, Zhang et al. (2019) [15] explored the application of deep learning to sentiment analysis, focusing on sentiment polarity. This research employed datasets like Sentiment140 and IMDB Reviews, utilizing techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings. Their comparative analysis revealed that Recurrent Neural Networks (RNN) and word embeddings achieved maximum accuracies of 0.82 on Sentiment140 and 0.87 on IMDB Reviews, respectively.

According to the research presented in the paper by Saravana Kumar et al.(2022)[9], all classifiers demonstrated good accuracy scores, except for the Ada Boost Classifier. Among the seven classifiers evaluated, Linear SVC, Perceptron, Passive Aggressive Classifier, and Logistic Regression emerged as the top performers. The classification was conducted using datasets related to COVID-19.

Aron Culotta (2010) [16] gathered tweets pertaining to influenza from Twitter. They conducted sentiment classification on these tweets using various machine-learning classifiers. The results indicated that multiple linear regression models outperformed the other classifiers.

In the research by Das Adhikary et al. (2022) [18], BERT achieved the highest F1-scores, scoring 0.9380 on the IMDB dataset and 0.8114 on the Sentiment 140 dataset, followed by GPT-3 with scores of 0.9119 and 0.7913, and Bi-LSTM with scores of 0.8971 and 0.7778 in the first stage. In the second stage, GPT-3 excelled in sentiment analysis of partially annotated COP9 conference-related tweets, attaining an F1-score of 0.8812. This study highlights the superiority of pre-trained models like BERT and GPT-3 for sentiment analysis tasks, surpassing traditional methods on standard datasets. Additionally, GPT-3's strong performance on partially annotated COP9 tweets underscores its capability to generalize well to domain-specific data with limited annotations, offering researchers and practitioners a viable solution for sentiment analysis in scenarios with scarce or absent annotated data across various domains.

Wahyu, Calvin, Frans, Mariel, Siti Mariyah, and Setia Pramana (2013) [13] conducted a study where they integrated various classification and feature extraction methods. Their research revealed that deep learning neural networks outperformed SVM and Naïve Bayes classifiers. Notably, their analysis indicated that employing a deep-learning neural network in conjunction with Bigram feature extraction yielded the most favorable outcomes.

III. METHODOLOGY

The methodology encompasses several key steps:

- Data Collection: Twitter tweets are collected using the sentiment 140 dataset, focusing on specific topics or hashtags relevant to the sentiment analysis task. The dataset comprises a mix of positive, and negative tweets, ensuring a diverse range of sentiment expressions.
- Preprocessing: The collected tweets undergo preprocessing to clean and standardize the text data. This includes tokenization, lowercasing, punctuation removal, stop-word removal, and stemming to enhance the quality of the text data and remove noise.
- Feature Extraction: Textual features are extracted from the preprocessed tweets using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency). These features serve as input to the sentiment analysis model.
- Model Selection and Training: Various machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, and Regression are explored for sentiment analysis. The selected model is trained on the labeled tweet dataset to learn the relationships between textual features and sentiment labels.
- Model Evaluation: The trained sentiment analysis model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques may be employed to assess the model's generalizability and robustness.

Assumptions Made

- The assumption is made that the labeled tweet dataset accurately represents the sentiment distribution in real-world Twitter data.
- It is assumed that the preprocessing steps effectively remove noise and irrelevant information from the tweet text, enhancing the quality of the input data for sentiment analysis.
- The selected machine learning or deep learning algorithms are assumed to be suitable for the task of sentiment classification and capable of capturing the complex relationships between textual features and sentiment labels.

3.1 Proposed Mechanism

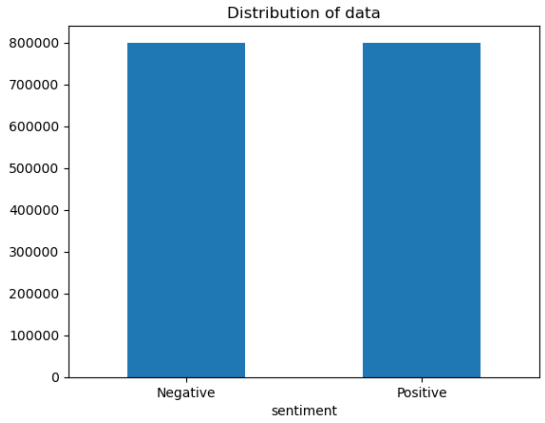

Figure 1: Distribution of data

We're creating 3 different types of models for our sentiment analysis problem:

- Bernoulli Naive Bayes
- Linear Support Vector Classification
- Logistic Regression

Since our dataset is not skewed, we're choosing Accuracy as our evaluation metric. Furthermore, we're plotting the Confusion Matrix to get an understanding of how our model is performing on both classification types.

3.2 Graphical / Tabular Form:

1. Accuracy and F1-Score Comparison:

- Bernoulli Naive Bayes

	precision	recall	F1-score	support
0	0.81	0.79	0.80	39989
1	0.80	0.81	0.81	40011
accuracy			0.8	80000
macro avg	0.80	0.80	0.80	80000
weighted avg	0.80	0.80	0.80	80000

- Linear Support Vector Classification

	precision	recall	F1-score	support
0	0.82	0.81	0.82	39989
1	0.81	0.83	0.82	40011
accuracy			0.8	80000
macro avg	0.82	0.82	0.82	80000
weighted avg	0.82	0.82	0.82	80000

- Logistic Regression

	precision	recall	F1-score	support
0	0.83	0.82	0.83	39989
1	0.82	0.84	0.83	40011
accuracy			0.83	80000
macro avg	0.83	0.83	0.83	80000
weighted avg	0.83	0.83	0.83	80000

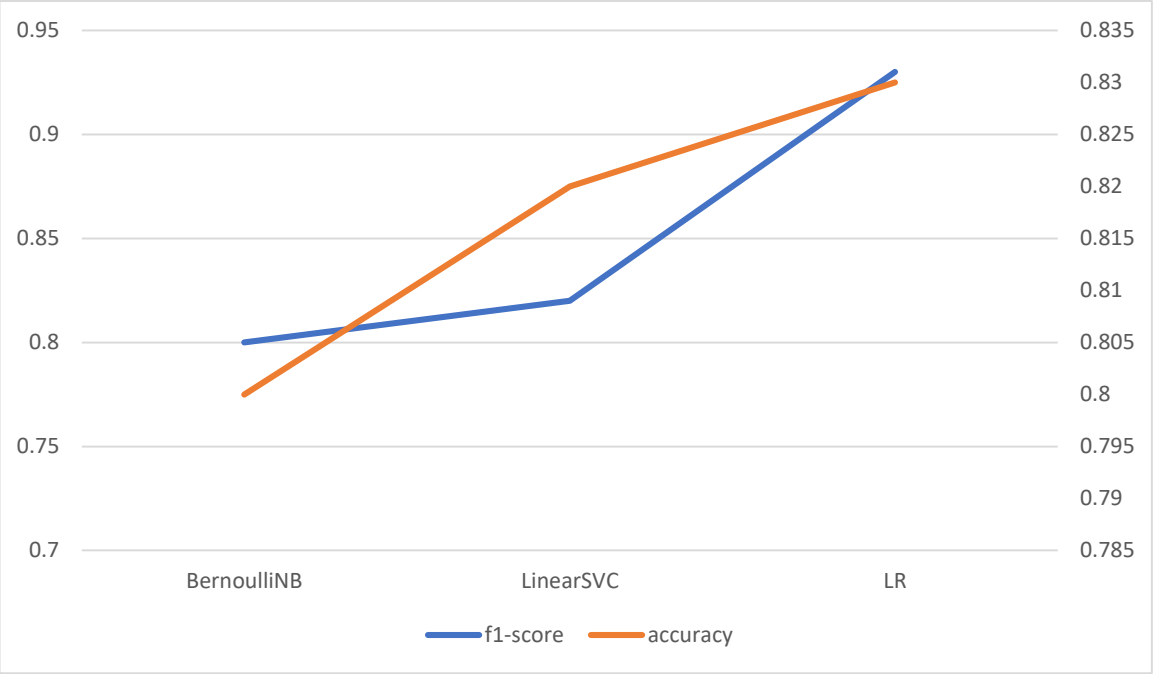

Figure 2: Graph representation of f1-score and accuracy

3.3 Confusion Matrix :

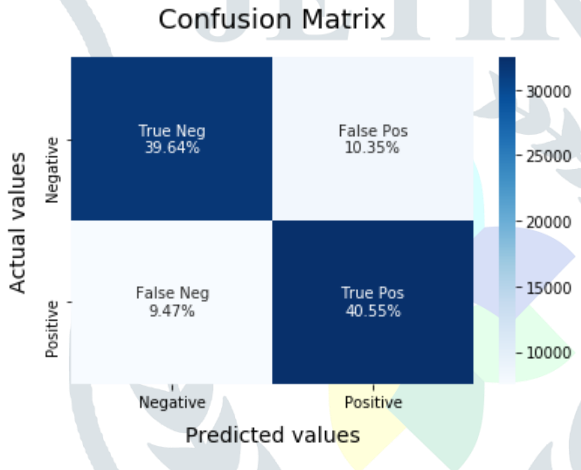

Figure 3: Confusion Matrix of Bernoulli Naive Bayes

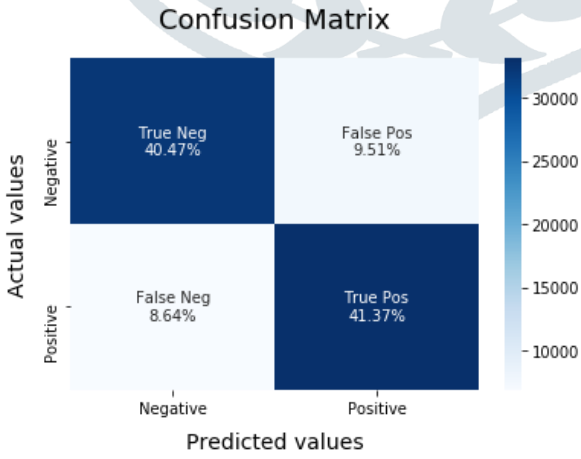

Figure 4: Confusion Matrix of Linear Support Vector Classification

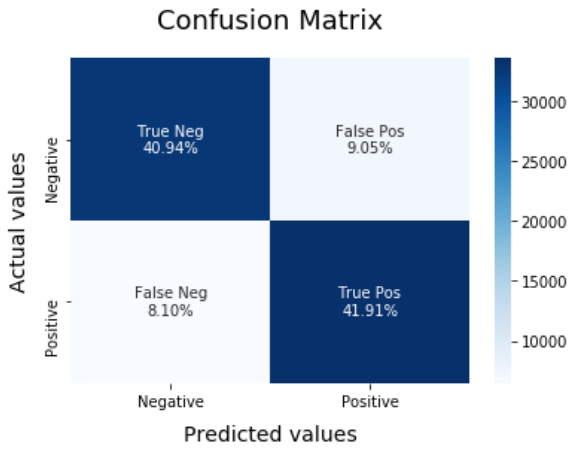

Figure 5: Confusion Matrix of Logistic Regression

IV. RESULTS AND DISCUSSION

The high accuracy and F1-score achieved by the model underscore the advancements in natural language processing and its applications in sentiment analysis. The results signify that it is highly effective at understanding and classifying sentiments in tweets. This has practical implications for industries relying on sentiment analysis for customer feedback, market research, and social media monitoring.

We can see that the Logistic Regression Model performs the best out of all the different models that we tried. It achieves nearly 82% accuracy while classifying the sentiment of a tweet.
However, it should also be noted that the BernoulliNB Model is the fastest to train and predict. It also achieves 80% accuracy while classifying.

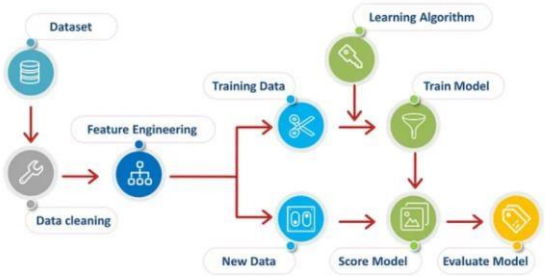

Figure 6: Different steps of data pre-processing[21]

V. CONCLUSION

The results of the sentiment analysis on the Sentiment140 dataset indicate that the selected models can effectively classify tweet sentiments with a high degree of accuracy. Among the various models tested demonstrated superior performance in capturing the nuances of tweet sentiments compared to traditional machine learning models. The methodology adopted for this research involved several key steps: data preprocessing, feature extraction, model training, and evaluation. Data preprocessing included cleaning the tweets by removing irrelevant characters, handling missing values, and normalizing text. Feature extraction was performed using methods such as TF-IDF (Term Frequency-Inverse Document Frequency). Various machine learning models, including logistic regression, support vector machines, and naive bayes, were trained and evaluated to identify the best-performing model. Performance metrics such as accuracy, precision, recall, and F1-score were used to evaluate the models. The preprocessing and feature extraction techniques played a crucial role in enhancing model performance by ensuring that the input data was clean and well-represented.

The significance of the results lies in their practical applications. Accurate sentiment analysis of tweets can provide valuable insights for businesses, policymakers, and researchers. For instance, companies can use sentiment analysis to gauge public opinion about their products and services in real time, allowing for timely interventions and improved customer engagement. Policymakers can monitor public sentiment on various issues, helping them to make informed decisions and respond to public concerns more effectively. Furthermore, the successful application of advanced NLP and machine learning techniques in this research demonstrates the potential for similar approaches to be applied to other types of social media data and text analysis tasks.

The future scope of work would be:

- Enhanced Data Preprocessing Techniques

Future work could focus on enhancing data preprocessing techniques to further improve model performance. This includes implementing more sophisticated text normalization methods, such as handling slang, abbreviations, and emoticons more effectively. Additionally, exploring techniques for context-aware preprocessing, which takes into account the context in which words are used, could lead to better sentiment classification.

- Integration with Real-Time Data Streams

Another potential area for future research is the integration of the sentiment analysis model with real-time data streams. This would involve developing systems that can process and analyze tweets as they are posted, providing real-time sentiment insights.

Such systems would require efficient data handling and processing capabilities, as well as robust models that can quickly adapt to new and evolving trends in language use on social media.

- Multi-Modal Sentiment Analysis

This involves combining textual data with other forms of data such as images, videos, and audio found in tweets to provide a more comprehensive sentiment analysis. Techniques such as deep learning models that can handle multi-modal inputs would be explored, potentially leading to richer and more accurate sentiment insights.

In summary, this research has successfully demonstrated the capability of machine learning models to perform sentiment analysis on Twitter data. Future work can build upon these foundations to develop more sophisticated, real-time, and multi-modal sentiment analysis systems, enhancing their applicability and impact in various domains.

REFERENCES

[1]. Dhanta, R., Sharma, H., Kumar, V. & Singh, H. O. (2023). Twitter sentimental analysis using machine learning. International Journal of Communication and Information Technology, 4(1), 71–83. DOI: 10.33545/2707661x.2023.v4.i1a.63.

[2]. Jacob, S. S. & Vijayakumar, R. (2021). Sentimental analysis over twitter data using clustering based machine learning algorithm. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-020-02771-9.

[3]. Wankhade, M., Rao, A. C., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, challenges. Artif. Intell. Rev. 55, 5731–5780. doi: 10.1007/s10462-022-10144-1

[4]. P. Khurana Batra, A. Saxena, Shruti & C. Goel. (2020). Election result prediction using twitter sentiments analysis. Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, pp. 182-185. DOI: 10.1109/PDGC50313.2020.9315789.

[5]. Dashrath Mahto, Subhash Chandra Yadav & Gotam Singh Lalotra. (2022). Sentiment prediction of textual data using hybrid convbidirectional-lstm model. Mobile Information Systems. https://doi.org/10.1155/2022/1068554.

[6]. K. S. Madhu, B. C. Reddy, C. Damarukanadhan, M. Polireddy & N. Ravinder. (2021). Real time sentimental analysis on twitter. 6th International Conference on Inventive Computation Technologies, Coimbatore, India, pp. 1030-1034. DOI: 10.1109/ICICT50816.2021.9358772.

[7]. Çilgin, C., Baş, M., Bilgehan, H. & Ünal, C. (2022). Twitter sentiment analysis during covid-19 outbreak with VADER. AJIT-e: Academic Journal of Information Technology, 13(49), 72–89. https://doi.org/10.5824/ajite.2022.02.001.x.

[8]. Al-Natour, S., and Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. Int. J. Inf. Manag. 54:102132. doi: 10.1016/j.ijinfomgt.2020.102132

[9]. Gulati, K., Saravana Kumar, S., Sarath Kumar Boddu, R., Sarvakar, K., Kumar Sharma, D., Nomani, M. Z. M., et al. (2022). Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to covid-19 pandemic. Mater. Today. 51, 38–41. doi: 10.1016/j.matpr.2021.04.364

[10]. S. Tiwari, "Social Media Sentiment Analysis On Twitter Datasets," pp. 925–927, 2020.

[11]. A. Goswami et al., "Sentiment Analysis of Statements on Social Media and Electronic Media Using Machine and Deep Learning Classifiers," Computational Intelligence and Neuroscience, vol. 2022, 2022, doi: 10.1155/2022/9194031.

[12]. R. S. Wadawadagi and V. B. Pagi, "Sentiment Analysis on Social Media," no. April, pp. 508–527,
2020, doi: 10.4018/978-1-5225-9643-1.ch024.

[13]. Mariel, Wahyu Calvin Frans, Siti Mariyah, and Setia Pramana. "Sentiment analysis: a comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text." In Journal of Physics: Conference Series, vol. 971, no. 1, p. 012049. IOP Publishing, 2018.

[14]. Kiruthika M., Sanjana Woona, Priyanka Giri, "Sentiment Analysis of Twitter Data", International Journal of Innovations in Engineering and Technology, 2016

[15]. Zhang, Y., Song, D., Zhang, P., Li, X., and Wang, P. (2019). A quantum-inspired sentiment representation model for Twitter sentiment analysis. Appl. Intell. 49, 3093–3108. doi: 10.1007/s10489-019-01441-4

[16]. Culotta, Aron. (2010). Towards detecting Influenza Epidemics by Analyzing Twitter Messages. SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics. 10.1145/1964858.1964874.

[17]. Cambria, E.; Mao, R.; Han, S.; Liu, Q. Sentic parser: A graph-based approach to concept extraction for sentiment analysis. InProceedings of the 2022 International Conference on Data Mining Workshops, Orlando, FL, USA, 30 November–3 December 2022.

[18]. Paul, B., Guchhait, S., Dey, T., and Das Adhikary, D. (2022). "A comparative study on sentiment analysis influencing word embedding using SVM and KNN," in Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021 (Springer Singapore), 199–211. doi: 10.1007/978-981-16-4284-5_18

[19]. N. Deepa, J. S. Priya & T. Devi. (2023). Sentimental analysis recognition in customer review using Novel-CNN. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1-4. doi: 10.1109/ICCCI56745.2023.10128627.

[20]. Rodrigues, A.P.; Fernandes, R.; Shetty, A.; Lakshmanna, K.; Shafi, R.M. Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Comput. Intell. Neurosci. 2022,2022, 5211949.

[21]. Kumari Anjali Suman and Vandana Bhattacharjee,"Heart disease Prediction – A Performance analysis of Machine Learning Classifiers", Journal of Emerging Technologies and Innovative Research (JETIR) May 2024, Volume 11, Issue 5