Sentiment Analysis Techniques: A Comparative Study of Logistic Regression, Random Forest, and Naive Bayes on General English and Nigerian Texts

Original Research Article

ABSTRACT

This research investigates sentiment analysis on two distinct datasets: a general English dataset and a Nigerian dataset (Gangs of Lagos movie review), using three machine learning algorithms: Logistic Regression, Random Forest, and Naive Bayeswith python programming language and its libraries. The study aims to evaluate and compare the performance of these models across different linguistic and cultural contexts. Results indicate that Logistic Regression consistently outperforms the other models, achieving the highest accuracy and balanced performance across sentiment classes. Random Forest provides comparable results but struggles with positive sentiment detection in the Nigerian dataset. Naive Bayes shows the lowest overall accuracy, with significant challenges in recall for certain sentiment classes. These findings highlight the importance of model selection and tuning tailored to specific datasets for effective sentiment analysis.

KEYWORDS

Sentiment Analysis, Machine Learning, Natural Language Processing, Naive Bayes, Supervised Learning, Artificial Intelligence

I. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a crucial field in natural language processing that involves determining and extracting emotions, opinions, and attitudes expressed

in text data. In this comprehensive survey, we explore the evolution and applications of sentiment analysis with a specific focus on the utilization of machine alearning techniques. Sentiment analysis aims to comprehend the sentiment or emotion conveyed in textual data, whether it's positive, negative, or neutral. The importance of sentiment analysis lies in its ability to provide valuable insights into public opinion, customer feedback, and social trends. Businesses leverage sentiment analysis to enhance decision-making processes, refine marketing strategies, and improve overall customer satisfaction (Prastyo et al., 2020).

Beyond business, sentiment analysis influences political landscapes, offering insights into public opinions about political figures, policies, and events. In social sciences, it becomes a lens through which emerging trends and societal shifts are observed. In market research, sentiment analysis aids in understanding consumer preferences and shaping targeted marketing campaigns (Nugroho, 2021). The advent of social media has amplified the role of sentiment analysis. Monitoring and analyzing conversations on platforms like Twitter, Facebook, and Instagram provide organizations with a dynamic understanding of audience sentiments (Karamollaoğlu et al., 2018). Sentiment Analysis is an intersection of Natural Language Processing and Machine learning as shown in the figure below:

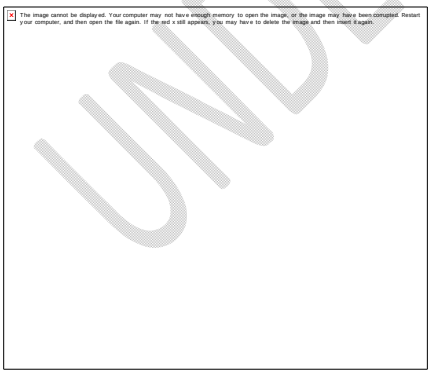

Figure 1. Sentiment Analysis

As we delve deeper into this research, the subsequent sections will unravel the intricacies of sentiment analysis methodologies, shedding light on the ongoing innovations and the role of machine learning in deciphering sentiments from diverse textual sources.

II. THEORETICAL ANALYSIS

Approaches in sentiment analysis encompass various methodologies and techniques employed to discern and classify sentiments expressed in textual data. Here, we explore the prominent approaches:

A. Lexicon-based

Lexicon-based approaches in sentiment analysis involve utilizing pre-built sentiment lexicons or dictionaries to determine the sentiment of a given text. A sentiment lexicon is a curated collection of words, where each word is assigned a sentiment score or label (e.g., positive, negative, neutral). These lexicons are often created manually or through automated methods, incorporating sentiment annotations. In lexicon-based approaches, each word in the text is matched against the entries in the sentiment lexicon. The lexicon provides sentiment scores or labels for each word, reflecting the perceived sentiment associated with that word (Waspodo et al., 2022).

B. Machine learning

Machine learning approaches in sentiment analysis utilize computational models to classify sentiments in textual content as these models learn patterns and relationships from data, allowing them to make predictions on unseen text (Nugroho, 2021). Supervised learning models, including Support Vector Machines (SVM), Naive Bayes, and neural networks, are commonly employed. These models require labeled datasets for training, where each text is associated with its corresponding sentiment label (positive, negative, or neutral). Once trained,

the model generalizes from this labeled data to predict sentiments in new, unseen text (Prastyo et al., 2020). Unsupervised learning methods, such as clustering and topic modeling, offer an alternative. These approaches do not rely on labeled data but instead aim to discover patterns or groupings within the text (Jain and Kaushal, 2018). Clusters of similar sentiments can emerge, providing insights into the overall sentiment landscape. Machine learning's role in sentiment analysis is transformative, enabling automated, context-aware, and scalable sentiment classification. As machine learning evolves, sentiment analysis stands to benefit from advancements that improve accuracy and applicability across diverse domains (Neshan, 2020).

1) Contextual Understanding:

Contextual understanding in the context of sentiment analysis refers to the ability of algorithms to interpret and comprehend the nuanced meaning of language within its given context. It goes beyond simplistic keyword-based approaches and aims to capture the subtleties, nuances, and complexities of language, making the sentiment analysis more accurate and reflective of real-world communication (Li et al, 2020).

2) Transfer Learning for Domain Adaptation:

Transfer learning, a machine learning paradigm, allows sentiment analysis models to leverage knowledge gained from one domain and apply it to another. This adaptation is particularly valuable when sentiment analysis needs to be performed in a specific domain with limited labeled data. In the traditional machine learning paradigm, models are trained and tested on data from the same domain. However, real-world scenarios often present challenges where labeled data is scarce in the target domain. Transfer learning addresses this by enabling models to draw on knowledge acquired from a source domain where labeled data is more abundant (Sharma et al., 2020). It enables models to leverage knowledge from a more general source

domain, enhancing their adaptability and performance across diverse and specialized contexts (B. Zhang et al., 2022).

3) Classification of machine learning algorithms

Machine learning algorithms can be broadly classified into three main categories based on their learning processes and goals:

I. Supervised Learning

In this type of supervised learning, the algorithm is trained on a labeled dataset where the output (or target) is known. The goal is to learn a mapping from input features to predefined classes. Classification algorithms are used when the output is a category or label, such as spam or not spam, disease or no disease (Prastyo et al., 2020). It can also involve predicting a continuous output variable. The algorithm learns a mapping from input features to a continuous output. Regression algorithms are used when the output is a quantity, like predicting house prices based on features like size and location (Sindhu et al., 2023).

II. Unsupervised Learning

Unsupervised learning involves algorithms that work with unlabeled data, aiming to find hidden patterns or structures within the dataset. Clustering algorithms group similar data points together based on certain criteria, without predefined classes. Common applications include customer segmentation, image segmentation, and anomaly detection (Jain and Kaushal et al., 2018)

III. Reinforcement Learning

Reinforcement learning involves an agent interacting with an environment and learning to make decisions to achieve a goal. The agent receives feedback in the form of rewards or

punishments based on its actions. The goal is to learn a policy that maximizes the cumulative reward over time. Applications include game playing, robotic control, and autonomous systems (Sharma et al., 2022).

4) Supervised Machine learning algorithms used in sentiment analysis

Machine learning algorithms are computational approaches designed to learn patterns and relationships from data, enabling them to make predictions or decisions without explicit programming. These algorithms form the backbone of various applications, including sentiment analysis. Let's explore the machine learning algorithms used in this research:

I. Logistic Regression

It calculates the probability of a sample belonging to a particular class and applies a logistic function to make predictions. In sentiment analysis, it can be utilized to predict the likelihood of a text being positive or negative (Cam et al., 2023).

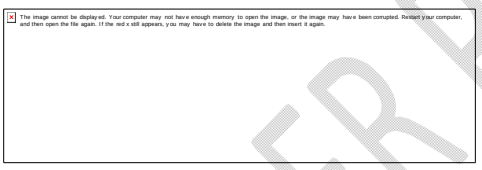

Figure 2. Logistic regression

II. Random Forest

Random Forests are an ensemble learning technique that combines multiple decision trees to improve overall accuracy and robustness. Each tree in the forest is trained on a different subset of the data and contributes to the final prediction (Umarani et al., 2021)

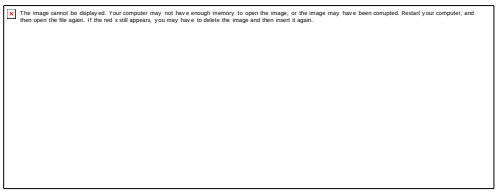

Figure 3. Random Forest

III. Naive Bayes

Based on Bayesian probability theory, Naive Bayes is a probabilistic classifier that assumes independence between features. It calculates the probability of a document belonging to a particular class given its features (Prastyo et al, 2020).

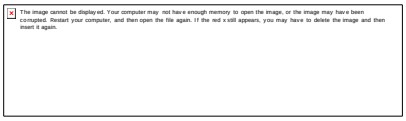

Figure4. Naive Bayes

These algorithms work by learning from labeled training data. During the training phase, they adjust their internal parameters based on the input data to minimize the difference between predicted and actual outcomes. Once trained, the model can generalize its knowledge to make predictions on new, unseen data The figures above are from their decisions on synthetic training data.

III. METHODOLOGY

The project involves using two datasets: a General English dataset and a Nigerian dataset called "Gangs of Lagos." The General English dataset consists of text data in English, representing a wide range of sentiments across various domains (27,480), including positive, negative, and neutral sentiments. This dataset serves as a baseline for sentiment analysis in standard English language contexts. The Nigerian dataset is specific to Nigerian English and Pidgin English, featuring text data from the "Gangs of Lagos" context (11,496 instances), and includes localized

sentiments reflecting cultural nuances and language variations unique to Nigeria. To manage and analyze the datasets, Python Pandas is utilized for data manipulation and analysis, providing functionalities such as DataFrames, data cleaning, data transformation, and exploratory data analysis. Scikit-learn, a machine learning library in Python, is used for model training, evaluation, data splitting, and feature extraction. It offers simple and efficient tools for data mining and data analysis, including classification, regression, and clustering. For visualization, Matplotlib, a comprehensive library for creating static, animated, and interactive visualizations in Python, is employed to plot graphs for data distribution, model performance, and result comparison.

The methodology begins with data preprocessing, which involves cleaning the text data by removing stop words, punctuation, and special characters, followed by tokenization. Pandas is used for data manipulation. The datasets are then split into training and testing sets using an 80:20 ratio, with Scikit-learn's `train_test_split` function ensuring a randomized and fair distribution of data. Three supervised machine learning models, such as Naive Bayes, Logistic Regression, and Random Forests, are used for sentiment classification.

Evaluation matrics and procedure

Evaluation metrics serve as the cornerstone of assessing the efficacy of sentiment analysis models in machine learning, offering a quantitative lens through which their performance can be comprehensively understood.

1). Confusion Matrix:

It is a tabular representation that breaks down the model's predictions into four categories: True Positives, False Positives, True Negatives, and False Negatives. It provides a detailed

understanding of the model's performance (Sapthami et al., 2023) as can be seen in the figure below.

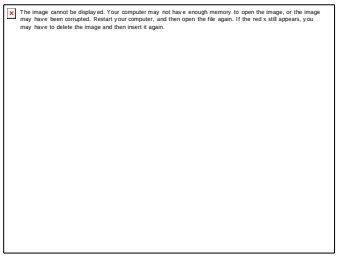

Figure5. Confusion Matrix

True Positive (TP):

Instances correctly predicted as positive. In sentiment analysis, it reflects the model's ability to accurately identify positive sentiments.

True Negative (TN):

Instances correctly predicted as negative. It represents the model's accuracy in identifying negative sentiments.

False Positive (FP):

Instances incorrectly predicted as positive. This indicates the model's tendency to make mistakes by identifying a positive sentiment when it's not present.

False Negative (FN):

Instances incorrectly predicted as negative. It shows cases where the model fails to identify a positive sentiment.

F1 Score: The harmonic mean of precision and recall. It provides a balanced measure, especially in situations with imbalanced datasets, where one sentiment class may dominate.

Recall (Sensitivity or True Positive Rate): Measures the ability of the model to correctly identify all relevant instances of a positive sentiment. High recall indicates that the model captures most positive sentiments.

Precision-Recall Curve (PR Curve): Visualizes the trade-off between precision and recall at different decision thresholds. It is particularly useful in scenarios where the dataset is imbalanced.

Precision: Measures the accuracy of positive predictions. High precision indicates fewer instances where the model incorrectly identifies a positive sentiment.

Accuracy: Measures the proportion of correct predictions (both positive and negative) out of all predictions made by the model. High accuracy indicates that the model correctly classifies most instances, regardless of whether they are positive or negative.

Metrics collectively offer a comprehensive view of a sentiment analysis model's strengths and weaknesses, helping in making informed decisions about its performance and potential areas for improvement (Umarani et al., 2021).

2) Procedure for sentiment analysis implementation

I. Input Text: Beginning with the raw text data as the input to the system.

II. Preprocessing: Removal of irrelevant characters, symbols, and special characters and handling uppercase/lowercase consistency.

III. Feature Extraction: Converting the preprocessed text into a format suitable for analysis and representing words as vectors capturing semantic relationships.

IV. Model selection : Choosing a sentiment analysis algorithm or model based on the nature of the problem and dataset.

V. Training: Training the selected machine learning model using a labeled dataset and adjusting of hyper-parameters for optimal performance.

VI. Testing and evaluation: Evaluating the model's performance on a separate set of labeled data (testing set) and Using metrics like accuracy, precision, recall, and F1 score.

VII. Decision Point - Accuracy Check: checking if the accuracy is acceptable.

VIII. Prediction (output): Applying the trained model to new, unseen text data to predict sentiments (Sapthami et al., 2023).

3) General Flowchart for sentiment analysis implementation

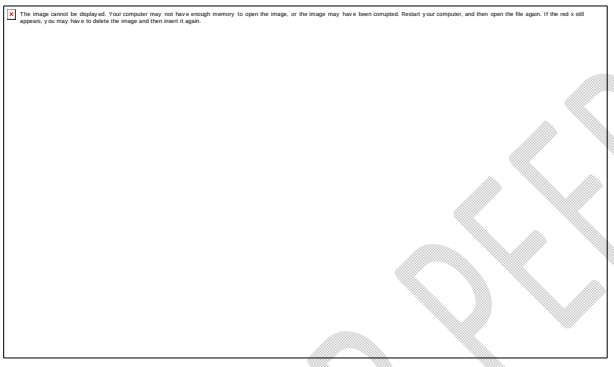

Figure 6. Generalized Flowchart

This flowchart above indicates the steps involved in the process using python programming language for the model creation and training.

IV. RESULTS AND DISCUSSION

The results, highlighting key metrics such as accuracy, precision, recall, F1-score and the tradeoff graphs between recall and precision are shown for each algorithm and dataset below:

1). For the General English Tweets

Classification Report - Logistic Regression:

	precision	recall	f1-score	support
neutral	0.81	0.78	0.80	1572
negative	0.79	0.88	0.83	2236
positive	0.91	0.80	0.85	1688
accuracy			0.83	5496

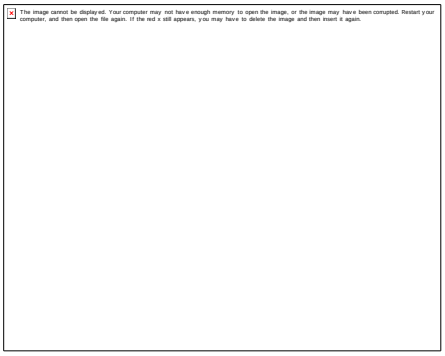

Image 1. Classification Report - Random Forest:

	precision	recall	f1-score	support
neutral	0.79	0.76	0.78	1572
negative	0.77	0.90	0.83	2236
positive	0.92	0.76	0.83	1688
accuracy			0.82	5496

Image 2. Classification Report - Naive Bayes:

	precision	recall	f1-score	support
neutral	0.92	0.58	0.71	1572
negative	0.68	0.96	0.79	2236
positive	0.92	0.73	0.81	1688
accuracy			0.78	5496

Image 3. For Gangs of Lagos Movie

Classification Report - Logistic Regression:

	precision	recall	f1-score	support
positive	0.90	0.44	0.59	399
neutral	0.85	0.94	0.89	1312
negative	0.85	0.89	0.87	1163
accuracy			0.85	2874

Image 4. Classification Report - Random Forest:

	precision	recall	f1-score	support
positive	0.94	0.20	0.33	399
neutral	0.82	0.89	0.85	1312
negative	0.74	0.87	0.80	1163
accuracy			0.79	2874

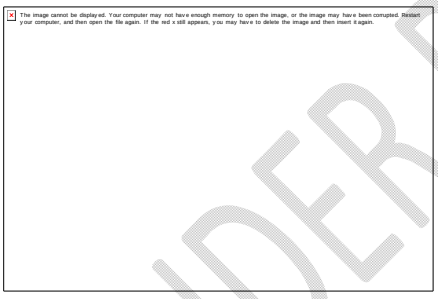

Image 5.Classification Report - Naive Bayes:

	precision	recall	f1-score	support
positive	0.86	0.03	0.06	399
neutral	0.88	0.76	0.81	1312
negative	0.63	0.94	0.75	1163

accuracy		0.73	2874

Image6.Precision Recall Curve

For the general English tweets, Logistic Regression achieved an accuracy of 83%. It performed strongly across all classes, with the highest F1-score for positive sentiments (0.85). Neutral and negative classes also showed robust performance, with F1-scores of 0.80 and 0.83, respectively. Random Forest had a slightly lower accuracy of 82%. Its performance was comparable to Logistic Regression, with consistent F1-scores across classes. Positive sentiments showed the highest precision (0.92) but had a lower recall (0.76). Naive Bayes, with an accuracy of 78%, had a lower overall performance compared to the other models. While it exhibited high precision for neutral and positive sentiments (0.92 each), it had a significantly lower recall for the neutral class (0.58), indicating many neutral tweets were misclassified.

For the Nigerian dataset focused on the "Gangs of Lagos" movie, Logistic Regression again demonstrated strong performance with an accuracy of 85%. It excelled in identifying neutral and negative sentiments, with F1-scores of 0.89 and 0.87, respectively. However, it had a lower recall for positive sentiments (0.44), suggesting difficulties in accurately identifying positive tweets. Random Forest had a lower accuracy of 79%, with weak performance in detecting positive sentiments (F1-score 0.33), though it handled neutral and negative sentiments reasonably well. Naive Bayes showed the lowest accuracy at 73%. It performed

poorly for positive sentiments (F1-score 0.06) but had decent precision and recall for neutral and negative sentiments, achieving the highest recall for the negative class (0.94).

V. CONCLUSION

In conclusion, this survey on sentiment analysis using machine learning algorithms has illuminated the dynamic landscape of extracting emotions, opinions, swings and attitudes from textual data. Overall, Logistic Regression consistently outperformed the other models in both datasets, excelling particularly in identifying positive sentiments in the general English dataset and neutral/negative sentiments in the Nigerian dataset. Random Forest provided balanced performance but struggled with positive sentiment detection in the Nigerian dataset. Naive Bayes, while showing the lowest overall accuracy, had significant recall issues in several categories, especially for positive sentiments in the Nigerian dataset. These results underscore the challenges and variances in sentiment analysis across different datasets, highlighting the importance of model selection and tuning for specific contexts.

It is evident that sentiment analysis, empowered by machine learning algorithms, is not just a technological tool; it is a catalyst for understanding the pulse of human sentiments in the vast sea of textual data. It is beneficial in business, political and security. The journey of sentiment analysis continues, marked by continual advancements and an unwavering commitment to unraveling the complexities of emotions embedded in the words we use. I recommend further sentiment analysis using more than three sentiment classes to capture the complexity in emotions.

Disclaimer (Artificial intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

REFERENCES

Assiroj, P., Kurnia, A. and Alam, S. (2023) 'The performance of Naïve Bayes, support vector machine, and logistic regression on Indonesia immigration sentiment analysis', Bulletin of Electrical Engineering and Informatics (BEEI), 12(1), pp. 235-240.

Cam, H., Cam, A. V., Demirel, U. and Ahmed, S. (2023) 'Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers', Heliyon. doi: 10.1016/j.heliyon.2023.e23784.

Gupta, S. (2024) 'What Is the Best Language for Machine Learning?', Springboard. Available at: https://www.springboard.com/blog/data-science/best-language-for-machine-learning (Accessed: March 2024).

Jain, K. and Kaushal, S. (2018) 'A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis', in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 483-487. doi: 10.1109/ICRITO.2018.8748793.

Karamollaoğlu, H., Doğru, I. A., Dörterler, M., Utku, A. and Yıldız, O. (2018) 'Sentiment Analysis on Turkish Social Media Shares through Lexicon Based Approach', in 2018 3rd

International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, pp. 45-49. doi: 10.1109/UBMK.2018.8566481.

Laeeq, F. and Tabrez, N. M. (2017) 'Sentimental Classification of Social Media using Data Mining', International Journal of Advanced Research in Computer Science, 8(5), pp. 546-549. doi: 10.26483/ijarcs.v8i5.3360.

Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., Liu, Q. and Xiang, T. (2020) 'Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis', IEEE Access, 8, pp. 46868-46876. doi: 10.1109/ACCESS.2020.2978511.

Masmoudi, A., Jamila, H. and Belguith, L. H. (2021) 'Deep Learning for Sentiment Analysis of Tunisian Dialect', Computacion y Sistemas. doi: 10.13053/cys-25-1-3472.

Neshan, S. A. S. and Akbari, R. (2020) 'A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis', in 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, pp. 8-14. doi: 10.1109/ICWR49608.2020.9122298.

Nugroho, D. K. (2021) 'US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis', in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 136-141. doi: 10.1109/Confluence51648.2021.9377201.

Pamukti, Y. B. P. and Rahardi, M. (2022) 'Sentiment Analysis of Bandung Tourist Destination Using Support Vector Machine and Naïve Bayes Algorithm', in 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, pp. 391-395. doi: 10.1109/ICITISEE57756.2022.10057802.

Prastyo, P. H., Sumi, A. S., Dian, A. W. and Permanasari, A. E. (2020) 'Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel', Journal of Information Systems Engineering and Business Intelligence, 6(2).

Sapthami, I., Krishna, B. M., Bhaskar, T. and Ravela, C. (2023) 'Sentiment Analysis using Machine Learning algorithms for Customer Product Reviews', in 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, pp. 447-451. doi: 10.1109/ICAISS58487.2023.10250522.

Sharma, P., Tomar, P. and Mukherjee, D. (2022) 'Sentiment Analysis on Amazon Dataset using Transfer learning', in 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP), Uttarakhand, India, pp. 160-165. doi: 10.1109/ICFIRTP56122.2022.10059413.

Sindhu, S., Kumar, S. and Noliya, A. (2023) 'A Review on Sentiment Analysis using Machine Learning', in 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, pp. 138-142. doi: 10.1109/ICIDCA56705.2023.10099665.

Umarani, V., Julian, A. and Deppa, J. (2021) 'Sentiment Analysis using various Machine Learning and Deep Learning Techniques', Journal of the Nigerian Society of Physical Sciences, 3(4), pp. 385-394.

Waspodo, B., Aini, Q., Singgih, F. R., Kusumaningtyas, R. H. and Fetrina, E. (2022) 'Support Vector Machine and Lexicon based Sentiment Analysis on Kartu Prakerja (Indonesia Pre-Employment Cards Government Initiatives)', in 2022 10th International Conference on Cyber

and IT Service Management (CITSM), Yogyakarta, Indonesia, pp. 01-06. doi: 10.1109/CITSM56380.2022.9935990.

Zaman, A. (2022) 'Public Sentiment Analysis of Covid-19 Vaccination Drive in Pakistan', Quaid-e-Awam University Research Journal of Engineering Science & Technology, 20(2), pp. 68-77.

Zhang, B., Chen, X., Ouyang, Y., Gan, Y., Lyu, B., Zhao, Q. and Li, C. (2022) 'Sentiment Analysis of Electric Power Domain based on Transfer Learning', in 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, pp. 738-742. doi: 10.1109/AEMCSE55572.2022.00148.