e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

COMPARATIVE ANALYSIS OF SENTIMENT ANALYSIS TECHNIQUES: SVM, LOGISTIC REGRESSION, AND TF-IDF FEATURE EXTRACTION

Hardik Jadia[*1]

[*1]UG Student, Department Of Information Technology, B. K. Birla College Kalyan,
(Empowered Autonomous Status), India.

DOI : https://www.doi.org/10.56726/IRJMETS45265

ABSTRACT

This research paper presents a comprehensive investigation into sentiment analysis using a diverse array of text classification models, focusing on classifying tweets into positive, negative, and neutral categories. The study encompasses multiple stages, commencing with data collection from online sources in CSV format, followed by rigorous data preprocessing. The text data undergoes feature extraction utilizing both CountVectorizer and TF-IDF Vectorizer techniques, facilitating logistic regression and Support Vector Machine (SVM) model training. The evaluation process employs a variety of metrics, including accuracy, precision, recall, F1-score, and ROC curves, to assess model performance. Three primary sentiment classification models are evaluated: Logistic Regression with CountVectorizer (lr_cv), Logistic Regression with TF-IDF Vectorizer (lr_tfidf), and Support Vector Machine (SVM). Each model exhibits unique characteristics in terms of precision, recall, and overall accuracy. Logistic Regression with CountVectorizer (lr_cv) achieves perfect recall for positive sentiments but at the expense of precision, resulting in misclassification of neutral and negative sentiments. Logistic Regression with TF-IDF Vectorizer (lr_tfidf) outperforms lr_cv, offering a balanced trade-off between precision and recall and a robust ROC curve. SVM emerges as the top-performing model with high accuracy, balanced precision, recall, and a strong ROC curve, demonstrating its efficacy in distinguishing between positive and negative sentiments. The choice of the most suitable sentiment analysis model depends on specific objectives, offering valuable insights for various applications. These findings contribute to informed model selection in the field of sentiment analysis, aiding researchers, practitioners, and decision-makers in choosing the most appropriate approach for their specific needs.

Keywords: Sentiment Analysis, Machine Learning Techniques, Support Vector Machine, Logistic Regression, Countvectorizer And TF-IDF Vectorizer Techniques.

I. INTRODUCTION

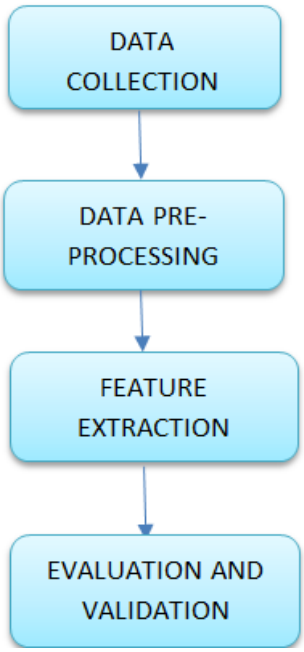
Sentiment analysis, a pivotal area of study, involves classifying chat messages as positive, negative, or neutral, extending to discerning underlying emotions such as anger, sadness, or happiness. This analytical process serves multifaceted purposes, aiding in mental health assessment and assisting businesses, developers, and manufacturers in gauging product satisfaction and profitability through online reviews.

II. LITERATURE REVIEW

Several notable studies have investigated the use of Support Vector Machines (SVM), Logistic Regression (LR), and TF-IDF feature extraction in the context of sentiment analysis, consistently yielding promising outcomes. SVM in Sentiment Analysis: Prior research has highlighted the effectiveness of Support Vector Machines in sentiment classification tasks. Notably, a study by [1] demonstrated the robustness of SVM models in capturing sentiment nuances, emphasizing their capacity to perform well in high-dimensional feature spaces. Logistic Regression for Sentiment Classification: Logistic Regression models have been explored extensively in sentiment analysis. A paper by [2] delves into the application of Logistic Regression for sentiment classification, showcasing the interpretability of this model while delivering competitive accuracy in distinguishing between positive, negative, and neutral sentiments.TF-IDF Feature Extraction: The significance of TF-IDF feature extraction in sentiment analysis is well-documented. [3] emphasized the role of TF-IDF in capturing term importance and rarity, which contributes to improved sentiment classification. Their study underscored the potential of TF-IDF as a valuable text feature extraction technique.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

III. METHODOLOGY



A. Data Collection:

Data for sentiment analysis was obtained from an online source in CSV format. The dataset includes various attributes such as tweet IDs, tweet text, timestamps, user information, and sentiment labels. The data is intended for sentiment analysis, with a focus on classifying tweets into different categories.

B. Data Pre-processing:

Data preprocessing involves loading a dataset, mapping sentiment labels to numerical values, cleaning the text by tokenization, punctuation removal, stopword removal, and lemmatization. The data is then split into training and testing sets, and text is vectorized using CountVectorizer and TfidfVectorizer. Logistic Regression (LR) and Support Vector Machine (SVM) models are trained and evaluated, and visualizations like word clouds, ROC curves, and confusion matrices are generated to assess model performance.

C. Feature Extraction:

The code employs two text feature extraction techniques, CountVectorizer and TF-IDF Vectorizer. CountVectorizer converts text data into a matrix where rows represent documents and columns represent unique words or n-grams, indicating word frequencies. TF-IDF Vectorizer enhances this by considering term importance based on both term frequency and its rarity across documents. These techniques are used with Logistic Regression for sentiment classification on Twitter data. The code includes data preprocessing, train-test splitting, model fitting using both techniques, and assessment through ROC curves and AUC scores for multi-class sentiment analysis.

D. Evaluation and validation:

Evaluation is performed through accuracy, precision, recall, F1-score, and ROC curves, providing insights into model performance. Grid search optimizes an SVM model, and key classification metrics, including ROC AUC, are reported. These steps ensure the robustness of sentiment analysis and model selection, facilitating reliable results for research purposes.

IV. MODELING AND ANALYSIS

TF-IDF and CF-IDF: Within the domain of text analysis, the Term Frequency-Inverse Document Frequency (TF-IDF) and Concept Frequency-Inverse Document Frequency (CF-IDF) techniques play a vital role. TF-IDF assesses word importance in document collections, while CF-IDF extends this concept to conceptual relevance. This paper examines their applications and implications in text analysis, contributing novel insights to the field.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

SVM: Support Vector Machines (SVMs) represent a powerful class of machine learning algorithms, widely employed in various fields. This research paper delves into the applications and advancements in SVMs, shedding light on their effectiveness in solving complex classification and regression problems, offering fresh perspectives and insights to the academic community.

LR: Logistic Regression (LR) is a fundamental statistical model extensively employed in diverse fields, including machine learning and social sciences. It serves as a predictive tool for binary classification tasks, offering a clear and interpretable framework. This research paper delves into LR's applications, methods, and potential advancements, contributing original insights to its utilization.

V. RESULTS AND DISCUSSION

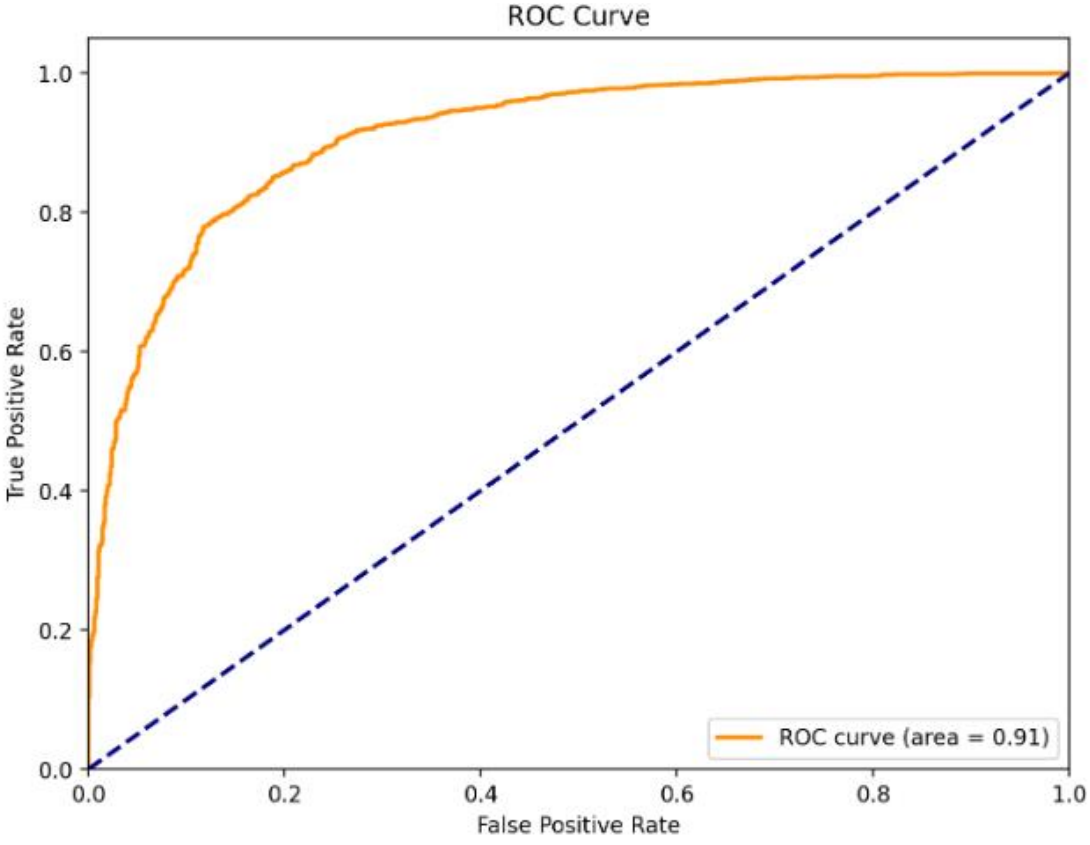
A. Accuracy:

Accuracy, a key metric, measures the model's overall correctness in classifying sentiments. It quantifies the proportion of correctly predicted sentiment labels and provides a fundamental assessment of the model's performance, with higher accuracy indicating better sentiment classification.

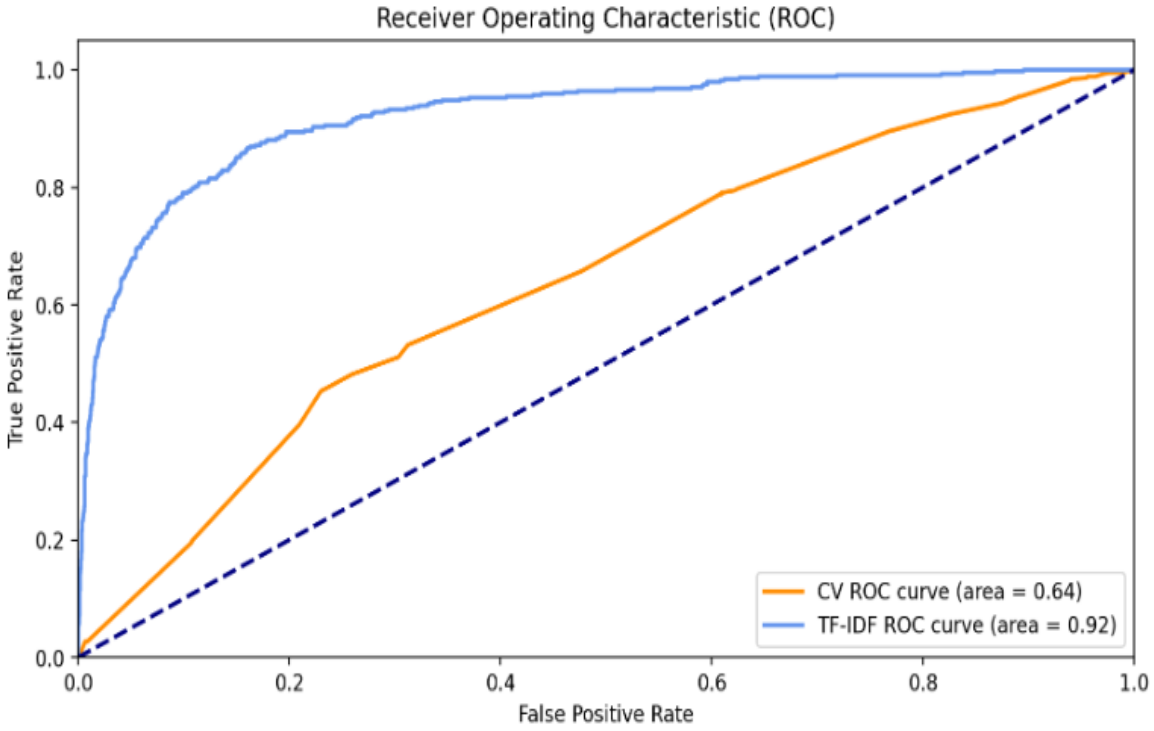
MODEL	ACCURACY	RECALL	PRECISION
Logistic Regression with CountVectorizer (lr_cv)	0.64	1.00	0.64
Logistic Regression with TF-IDF Vectorizer (lr_tfidf)	0.78	0.93	0.80
Support Vector Machine (SVM)	0.83	0.83	0.89

B. ROC Curve:

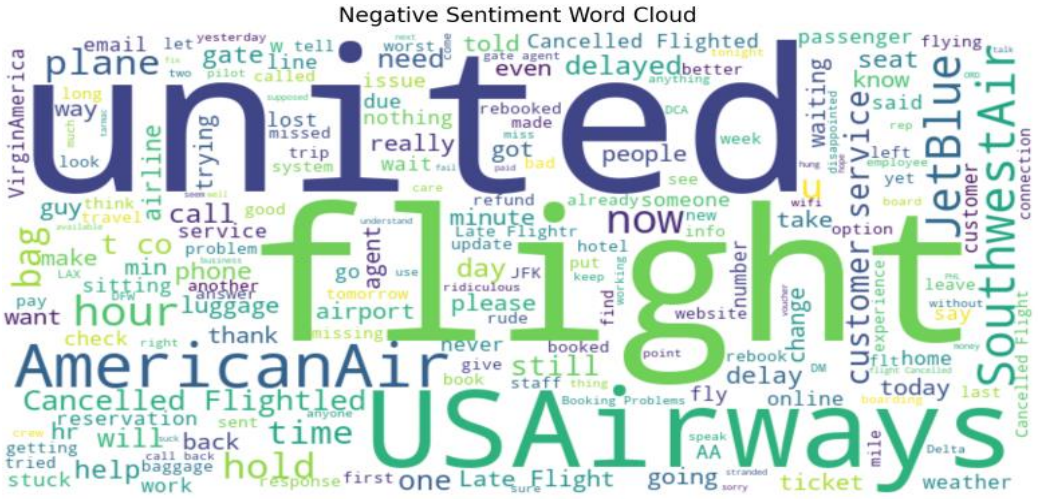
ROC curves and their corresponding AUC values are employed to assess the sentiment analysis model's ability to distinguish between sentiment categories. Higher AUC indicates better classification. ROC curves provide a visual representation of model performance, aiding readers in gauging the model's discriminatory power.

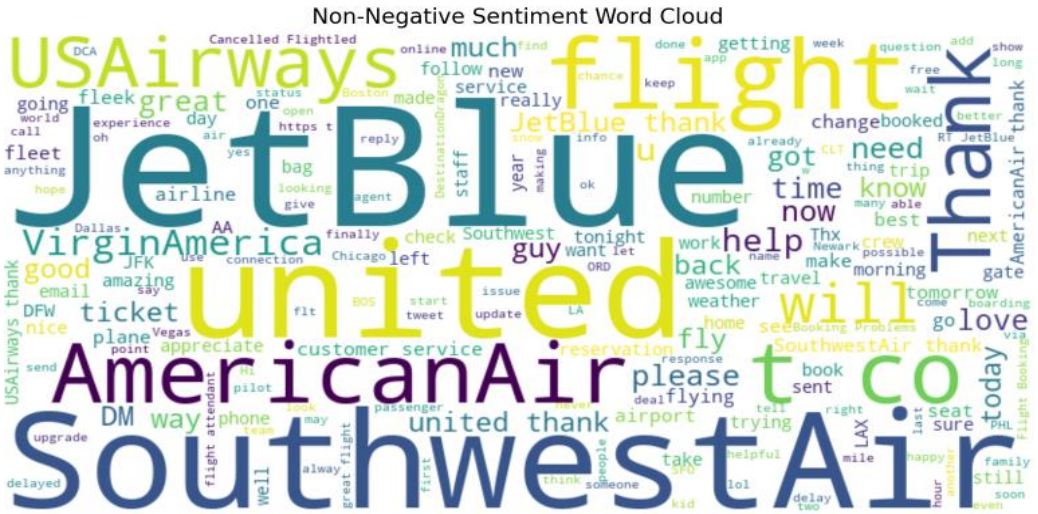


e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com



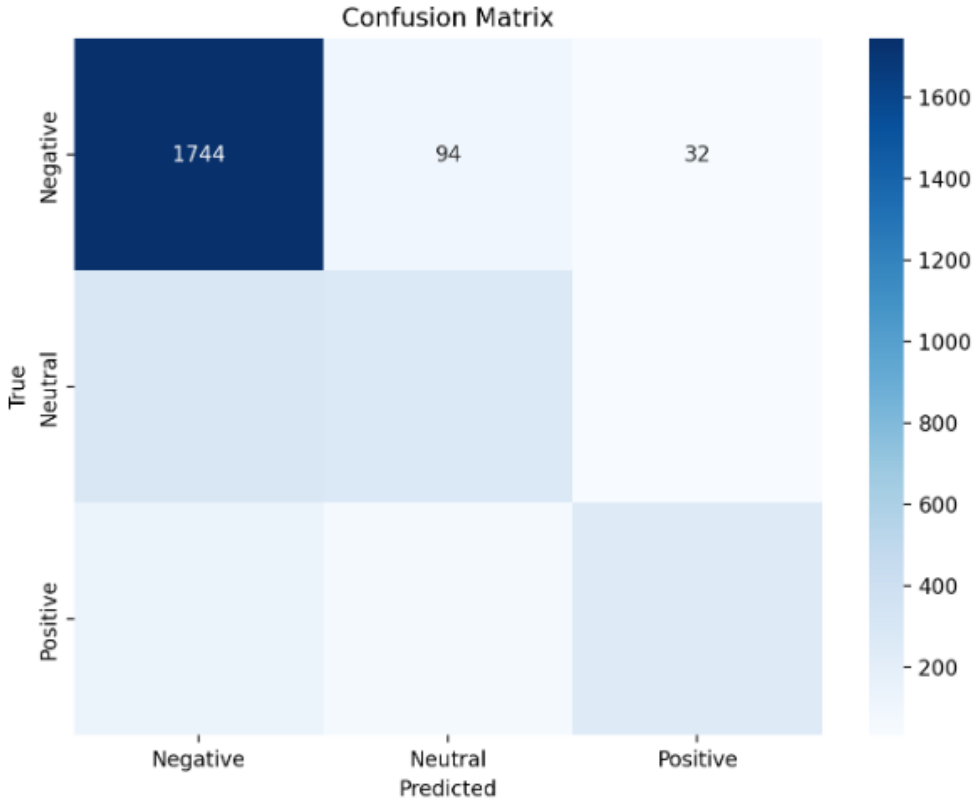
C. Word cloud:





e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

D. Confusion Matrix:



VI. CONCLUSION

In the context of sentiment analysis, three distinct models were evaluated: Logistic Regression with CountVectorizer (lr_cv), Logistic Regression with TF-IDF Vectorizer (lr_tfidf), and Support Vector Machine (SVM). Each model exhibited unique performance characteristics.

The lr_cv model demonstrated a perfect recall, effectively capturing all positive sentiments, but its precision was comparatively low, resulting in a notable number of false positives. This model's sensitivity to positive sentiment came at the expense of misclassifying some neutral or negative sentiments.

On the other hand, lr_tfidf outperformed lr_cv, boasting a higher accuracy, well-balanced precision and recall, and a robust ROC curve value of 0.92. This model offered a favorable trade-off between precision and recall.

SVM emerged as the top-performing model with an accuracy of 83% and a precision of 89%. Its balanced recall and high ROC curve value of 0.91 indicated strong discriminatory power between positive and negative sentiments.

The choice of the best model hinges on specific objectives. opt for lr_cv if prioritizing capturing as many positive sentiments as possible, even at the cost of some false positives. For a balanced model with a good precision-recall equilibrium, lr_tfidf is a compelling option. If precision is paramount and the goal is to minimize false positives, SVM stands as the preferred choice. These findings offer valuable insights into selecting the most suitable sentiment analysis model based on the specific requirements and trade-offs for a given application.

VII. REFERENCES

[1] Ahmad, Munir, Shabib Aftab, and Iftikhar Ali. "Sentiment analysis of tweets using svm." Int. J. Comput. Appl 177.5 (2017): 25-29.

[2] Yassine Al-Amrani, Mohamed Lazaar, and Kamal Eddine Elkadiri. 2017. Sentiment Analysis using supervised classification algorithms. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (BDCA'17). Association for Computing Machinery, New York, NY, USA, Article 61, 1–8. https://doi.org/10.1145/3090354.3090417

[3] V. Sundaram, S. Ahmed, S. A. Muqtadeer and R. Ravinder Reddy, "Emotion Analysis in Text using TF-IDF," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 292-297, doi: 10.1109/Confluence51648.2021.9377159.

[4] M. T. H. K. Tusar and M. T. Islam, "A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data," 2021 International Conference on Electronics, Communications, and Information Technology (ICECIT), Khulna, Bangladesh, 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641336.

[5] E. Aydoğan and M. A. Akcayol, "A comprehensive survey for sentiment analysis tasks using machine learning techniques," 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), Sinaia, Romania, 2016, pp. 1-7, doi: 10.1109/INISTA.2016.7571856.