Explainable Machine Learning for Twitter Sentiment Analysis: Understanding Why Tweets are Classified as Positive or Negative

Saed. S. Alewaity
Faculty of Information Technology, The Islamic University of Gaza (IUG), Gaza, Palestine

Abstract Today, we live in big data, the increase of data in the amount of user-generated data on social media platforms like Twitter [2] and more alot number of tweets every second, enabling individuals, brand organizations, that appeared on new challenges for companies in keep an eye on customer reviews and opinions about their products. while the traditional analytics tools try to improved and development to support larger datasets but need modern tools like Spark to handle big data.
This research paper combines Apache Spark because it's more flexible, fast, and scalable, machine learning techniques such as Logistic Regression and Random Forest, and data visualization methods to understanding why tweets classify as positive or negative in Twitter data, The system evaluation metrics using classification metrics such as accuracy, precision, recall, and F1 score and computational metrics like time efficiency, speed up. Experimental results showed Logistic Regression achieves higher accuracy and approximately 24× faster processing data compared to Random Forest, these results demonstrate that Logistic Regression is more suitable for large-scale Twitter sentiment and understanding tweets classify tasks and more time-efficient.

I. Introduction

NOWADAYS the with social media application like twitter, and have been trending in these days, they help people attitude or concern towards a certain topic. Users send messages (a.k.a., tweets) to a network of contacts from a wide variety of devices. A tweet is a text-based post and only has 140 characters, which is approximately the length of a typical newspaper headline and subhead [3]. Twitter [2] is a "what's- happening-right-now" social network and hence tweets are valuable sources for businesses, government and individuals to determine public's opinion or sentiment about an entity (product, people, topic, event etc). Understanding sentiment on social media platforms like twitter facilitates to company, governments and organizations to make decisions on time. But, the volume of tweets produced by Twitter every day is very vast (21 million tweets per hour, as measured in 2015) [4]. Hence there needed to automated the process of sentiment analysis without to read millions of tweets manually. Thats is speed up to understand of data, which allow to take fast decisions such as knowing customer opinion if positive or negative. Since social media sites handle huge amounts of data on a regular basis a powerful tool is required to manage, handle and retrieve it. Big Data would aid in achieving this. Big Data provide Apache Spark for data processing that can quickly perform processing tasks on very large data sets. Spark also takes some of the programming burdens of these tasks off the shoulders of developers with an easy-to-use API that abstracts away much of the grunt work of distributed computing and big data processing [5]. The main challenge is to process and analyze millions of tweets to understand why tweets classify as positive or negative which can't to be done manually or traditional tools. This paper aims to make a Sentiment Analysis and understand tweets classify application, we will use the Sentiment140 dataset with 1.6 million tweets from kaggle[6]. and spark Apache framework for data processing. train machine learning models like Logistic Regression and Random Forest. Use evaluation metrics using classification metrics such as accuracy, precision, recall, and F1 score and computational metrics like time efficiency, speed up.

II. LITERATURE REVIEW

In previous years, studies of Sentiment Analysis have wide attention. Because basically big size of data on social media especially that describe peoples, ideas and comments. And become Very important for Government, Organization and company.

Kumar et al. [10] (2023) proposed framework can be scalable and development for sentiment and analyzing large scale twitter data using Apache spark. The approach focused to improve to extract feature for enhance accuracy of sentiment classification. The system used Spark MLlib library for distributed processing. they demonstrated that optimizing extract feature improved both classification accuracy and time efficiency compared with traditional methods-based MapReduce. However, it not addresses explainability for reason to understanding why tweets classify as positive or negative. this limitation one of the gap our research aims to fill.

Shah and Malu [11] (2024) implemented a system of

scalable sentiment analysis system for twitter data using Apache spark to handle tweet in real time streaming. this study focused to overcoming challenges like big of size tweet data, unstructred data and short length of message. they applyed Naive Bayes, Logistic Regression, and Decision Tree algorithms after preprocessing for data like remove URLs, special characters, mentions, hashtag, extra spaces and stopwords. The results show Logistic Regression and Naive Bayes achieved higher accurcy compared to Decision Trees especially datasets size increased. and increasing number of nodes in spark cluster leads to reduce time processing significantly and improving efficiency. this study highlights to strong of spark in scalability and time efficiency for real time big data processing, but it not addresses explainability for reason to understanding why tweets classify as positive or negative. this limitation one of the gap our research aims to fill.

Elzayady, Badran, and Salama (2018) [13] presented framework of sentiment analysis twitter data using Apache spark. their approach dependand processing data Spark distributed processing and MLlib library to handle a lot of size tweets data efficiancy. They implemented preprocessing data with steps like remove URLs, special symbols, mentions, hashtag, extra spaces, stopwords and convert textt to lowercase. the evalution three classifcation algorithm: Naive Bayes, Logistic Regression, and Decision Trees on datasets have approximately 1,578,627 labeled tweets. Datasets split to multi sub sets for evalute scalability, and using train (70%) / test (30%). The results showed Naive Bayes and Logistic Regression achieved higher F-measure values compared to Decision Trees, especially dataset size increased. Additionally, found that adding more nodes to spark cluster reduced runing time thats proves scalability. but it not addresses explainability for reason to understanding why tweets classify as positive or negative. this limitation one of the gap our research aims to fill.

Baltas, Kanavos, and Tsakalidis (2017) [14] present implemetation for sentiment analysis system for twitter data using Apache spark to face challenges by data size, diversity and real time processing. They using Spark MLlib library with NLP processing techniques. They implemented steps preprocessing data. The system supported both binary and ternary sentiment calssification, and study how changes of size data set effected to classify quality. The aims study how scaling dataset influences effected to accuracy and runtime models on Spark. However, but it not addresses explainability for reason to understanding why tweets classify as positive or negative. this limitation one of the gap our research aims to fill.

Ahmed and Khan (2025) [15] present framwork of sentiment analysis using Random Forest algorithm to classify tweets real time as positive, neutral or negative. Dataset have 75681 inputs and 74995 unique tweet text. they implemented steps of preprocessing data including tokenization and lemmatization to prepare data for classify. Also publish web app based Stremlit for real time active with users. Their results show high accuracy for classification, showcasing effectivity for Random Forest with sentiment tasks cross three classes. This study focused primarily on classification performance, but it not addresses explainability for reason to understanding why tweets classify as positive or negative. this limitation one of the gap our research aims to fill.

Abhineswari and Priyadarshini [1] proposed a framework for real-time sentiment analysis of large-scale Twitter data using Apache Spark Streaming. The approach focused to handling high speed and size for tweets getting every second using big data processing. applyed multiple machine learning algorithms, like Naive Bayes, Random Forest, and Support Vector Machines (SVM), to classify the tweets. This study showed spark to help fast and scalability analysis compared to the traditional system. Experimental results showed that kernel-based methods such as Random Forest and SVM outperformed simpler classifiers in terms of accuracy, precision, recall, and F1-score. This work explains the effectiveness of combining Spark Streaming with manifold ML algorithms to enable real-time sentiment analysis over large-scale social media streams.

Jain and Dandannavar [4] (2016) proposed a more detailed apporched for sentiment analysis on twitter data machine learning techniques. their focused to classify tweets to category like positive, negative or natural with applies multinomial Naive Bayes and decision tree classifiers. this work underscores the importance of applying scalable machine learning models for handle big data from text generated like tweets.

Gupta, S., et al. [7] (2021) presented a comparative study on twitter sentiment analysis using machine learning and Hadoop system. their focused for challenges posed twitter data like slang, misspellings and brevity due to the character limit. they compare approach-based machine learning to extract sentiment. this work underscores the importance of scalable solutions when dealing with large and noisy data.

Krishnan et al. [8] (2021) performed a study on sentiment analysis for COVID-19 related with twitter data using machine learning models. this study foucesed classify the tweets to positive, negative, or neutral category with challenges for short text, informal language and huge data from social media posts. and applied Logistic Regression, Random Forest, and naive bayes algorithms for analysis public opinion in real time. this work emphasizes the importance of scalable frameworks for big data processing in social media analytics.

Saketh et al. [9] (2020) developed real time sentiment analysis system foe twitter data using spark. their system has interface to stream live tweets an applied the multinomial naive bayes classify for analyze sentiment. their hosted-on AWS providing friendly interface for analyze sentiment in real time. this study showed he effectiveness of apache spark in handling large data and sentiment analysis.

While previous studies focused to improving classification accuracy and scalability, but no studies have explainability. This paper fills this gap with provide system not only classify tweets but also understanding why classification tweet as positive or negative.

III. METHODOLOGY

This propose system aim to providing explainability to understand why tweets was classified positive or negative. The implemented workflow using python code with apache spark is a distributed computing framework designed for processing big datasets efficiently. This full source code is available on GitHub [16]

Main Stages:

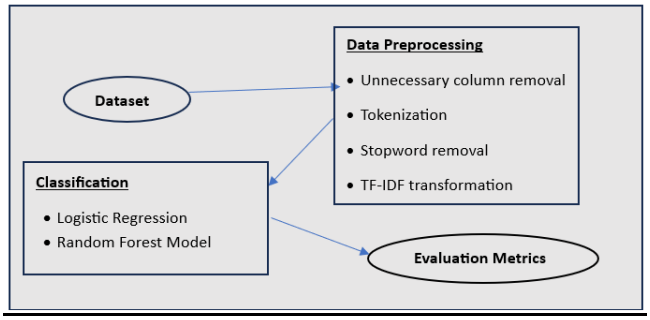

Fig. 1.0. Dataset diagram from preprocessing to evaluation.

The Fig. 1.0. show different of stages in system which include dataset, data preprocessing, models and evaluation metrics.

A. Data Preprocessing and Feature Engineering

Dataset was cleaned with steps mentioned above and transformed to structed format suitable for machine learning.

B. Model Training and Evaluation

We trained and compared two machine learning algorithms for text classification:

- Logistic Regression: Linear model suited for high dimensional sparse data like TF-IDF, and have high accuracy, fast training and interpretability through coefficients.
- Random Forest: Ensemble learning method based on decision trees, parameters (numTress=100, maxDepth=10, seed=42), have ability to capture non- linear relationship in data.

Training and Testing: Split the dataset for test (25%) and training (75%) and applied evaluation metrics with performance using classification metrics like accuracy, precision, recall, and F1-score and computational metrics like time efficiency, speed up.

C. Explainability Layer

In this stage provide to interpretability for decision using models.

- Logistic Regression:

Influence of each token: In positive what's word pushing to positive sentiment, in negative what's word pushing to negative sentiment. we defined top 20 positive (Fig. 2.0.) and top 20 negative tokens (Fig. 3.0.).

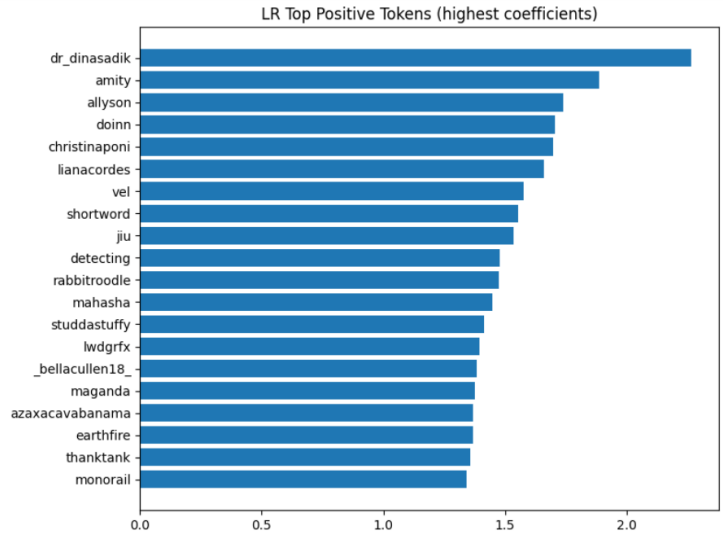

Fig. 2.0. Top positive tokens with Logistic Regression.

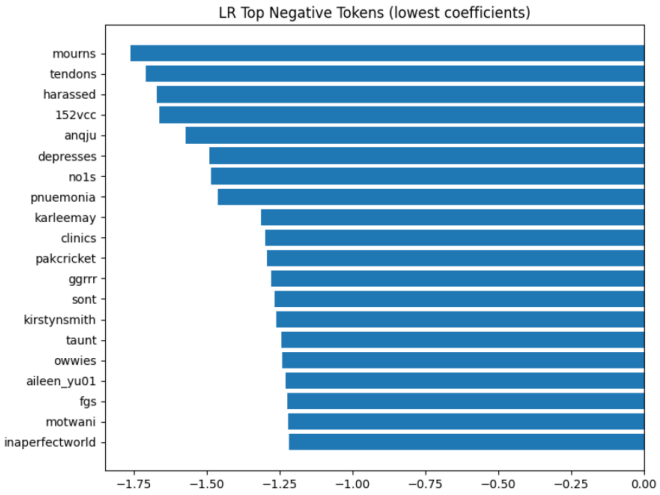

Fig. 3.0. Top Negative tokens with Logistic Regression.

- Random Forest:
  using "featureImportances" feature to extract overall influential word Fig. 4.0. To understanding why tweet was classify, Compute contribution every token in tweet:

$$\text{Contribution} = \frac{\text{TFIDF(Token)}}{\text{Coefficient (Token)}}$$

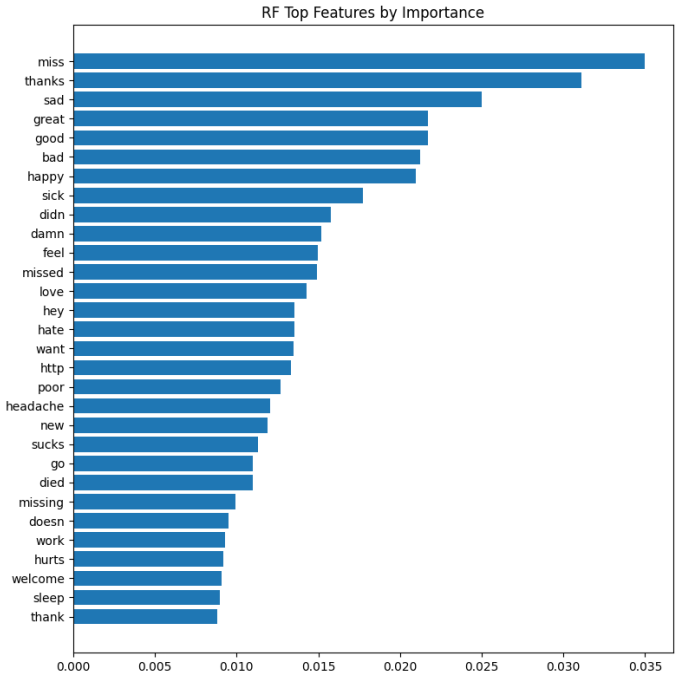

Fig. 4.0. Random Forest Top features by importance using (featureImportances feature).

IV. Experimental Setup

Experiments were conducted on Google Colab and Apache Spark 3.5.0. Sentiment140 dataset of 1.6 million tweets used, and included:

- Preprocessing text tweets
- Unnecessary column removal
- Tokenization
- Stopword removal
- TF-IDF transformation

Experiments were conducted on Google Colab and Apache Spark 3.5.0. Sentiment140 dataset of 1.6 million tweets used, and included:

- Accuracy
- Precision
- Recall
- F1-score
- Time efficiency

Speedup was calculated using:

$$\text{SpeedUp LR vs RF} = \frac{\text{RF time training}}{2a\text{LR time training}}$$

V. Results and Discussing

this section presents the experimental results of the understand why tweets classify positive or negative. Results focused on performance of two machine learning models: logistic Regression and Random Forest in terms of classification quality and efficiency.

A. Classification Metrics

Table [1] displayed the evaluation metrics for both models, including Accuracy, precision, Recall, and F1-score, time training (efficiency) and speedup.

Metrics / Model	Logistic Regression	Random Forest
Accuracy	0.778849	0.716442
Precision	0.779159	0.720500
Recall	0.778849	0.716442
F1-score	0.778782	0.715094
Training Time(efficiency)/ sec	285.27	6839.18
SpeedUp	X24	

Table 1.0. The models metrics Comparison.

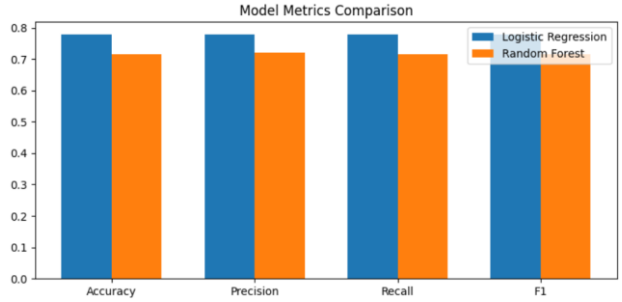

Fig. 5.0. The models metrics Comparison (Accuracy, Precision, Recall, F1-score).

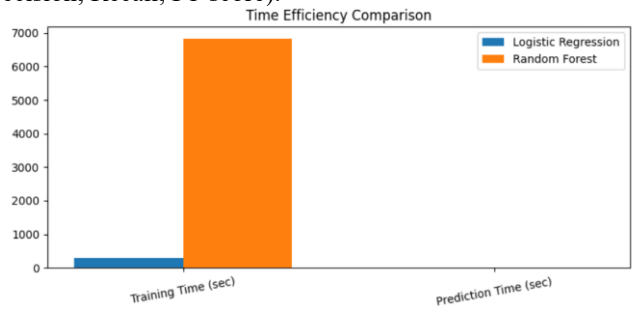

Fig. 6.0. The time efficiency comparison (Training time and prediction time).

Analysis:

Dependence on result in Table 1.0. Logistic Regression achieved higher accuracy compared to Random Forest with a 6.2% improvement. F1-score, which balances precision and recall, was also superior for Logistic Regression, indicating more consistent performance across positive and negative classes. Random Forest underperformed on textual data. Logistic Regression completed training in 285sec while Random Forest required 6839sec, Logistic Regression was approximately 24× faster compared to Random Forest, these results demonstrate that Logistic Regression is more suitable for large-scale Twitter sentiment classification tasks and more time-efficient.

VI. Conclusion

In this research, we built an interpretable machine learning pipeline for Twitter sentiment analysis at the scale of social data using Apache Spark.

The system that was proposed did not only categorize tweets to be positive or negative, but it also justified each message's evaluation by the classifier for transparency and trust in the model's opinions.
We experimented our approach with the Sentiment140 datasets that contains 1.6 million tweets. We fitted and compared two machine learning approaches such as: Logistic Regression and Random Forest.
We considered traditional metrics of classification Accuracy, Precision, Recall, and F1-score as well as computational score Training time and Prediction time that are vital for execution in Big Data impedance processors.
The results showed that:
Logistic Regression managed to surpass the performance of Random Forest in terms of both classification quality and running time.
The highest accuracy of the models was obtained using Logistic Regression (77.8%) which is higher than that reached by RF (71.6%).
The training of Logistic Regression was 24× faster than Random Forest, and its prediction was 7× faster, making LR a clear choice for scalable sentiment analysis tasks.
In addition performance, view the most inluential positive or negative words in dataset and show why tweet classified positive or negative. Despite success the system but this study faced limitation, it's focused on binary classify and English only data also limitation on testing environment, In future work we will supported multi language like Arabic text and work with deep learning models in spark.

ACKNOWLEDGMENT

The author would like to express his sincere gratitude to **Dr. Rebhi S. Baraka** for his supervision and guidance throughout this project, which was carried out as part of the requirements for the course **Big Data (ICTS 6339), Faculty of Information Technology, The Islamic University of Gaza (IUG).**

REFERENCES

[1] Abhineswari, M., & Priyadarshini, R. (2023). Analyzing Large-Scale Twitter Real-Time Streaming Data with Manifold Machine Learning Algorithms in Apache Spark.
[2] Twitter. (n.d.). Twitter. Retrieved August 13, 2025, from https://twitter.com
[3] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas. "Twitter and the micro-messaging revolution: Communication, connections" An OReilly Radar Report. 54 pages, November 2008.
[4] Jain, A. P., & Dandannavar, P. (2016). Application of machine learning techniques to sentiment analysis. In Proceedings of the International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 628–632). IEEE. https://doi.org/10.1109/ICATCCT.2016.7912076
[5] Saketh, V. S., Guntupalli, Y. K., & Vaishnav, D. S. (2020). Sentiment Analysis of Live Twitter Data using Apache Spark. International Research Journal of Engineering and Technology (IRJET), 7(8), 4873–4880.
[6] Kazanova, M. M. (2017). Sentiment140 dataset with 1.6 million tweets. Kaggle. https://www.kaggle.com/datasets/kazanova/sentiment140
[7] Gupta, S., Goswami, R. S., & Choubey, A. K. (2021). Twitter sentiment analysis using Machine Learning and Hadoop: A comparative study. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 680–685. https://doi.org/10.1109/Confluence51648.2021.9478077
[8] Krishnan, H., et al. (2021). Machine learning-based sentiment analysis of coronavirus disease-related Twitter data. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 686–691. https://doi.org/10.1109/Confluence51648.2021.9478145
[9] Saketh, V. S., Guntupalli, Y. K., & Vaishnav, D. S. (2020). Sentiment analysis of live Twitter data using Apache Spark. International Research Journal of Engineering and Technology (IRJET), 7(8), 836–840. Retrieved from https://www.irjet.net/archives/V7/i8/IRJETV7I8836.pdf
[10] Kumar, M., Verma, S., & Vashisth, S. (2023). A Feature Extraction based Improved Sentiment Analysis on Apache Spark for Real-time Twitter Data. Scalable Computing: Practice and Experience, 24(4), 305–316. https://doi.org/10.12694/scpe.v24i4.2343
[11] Shah, P., & Malu, N. (2024). An Apache Spark implementation for sentiment analysis on Twitter data. International Research Journal of Modernization in Engineering Technology and Science, 6(3), 1857-1861
[12] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf.Retrieval 2(1–2), 1–135 (2008) .
[13] Elzayady, H., Badran, K., & Salama, G. I. (2018). Sentiment Analysis on Twitter Data using Apache Spark Framework. In 2018 13th International Conference on Computer Engineering and Systems (ICCES). IEEE. doi:10.1109/ICCES.2018.8639195
[14] Baltas, A., Kanavos, A., & Tsakalidis, A. (2017). An Apache Spark implementation for sentiment analysis on Twitter data. International Journal of Computer Applications, 175(4), 7-14. https://www.researchgate.net/publication/315913579_An_Apache_Spark_Implementation_for_Sentiment_Analysis_on_Twitter_Data
[15] Ahmed, Z., & Khan, D. (2025). A Random Forest Approach For Real-Time Sentiment Analysis Of Twitter Data. Journal of Emerging Trends and Novel Research (JETNR), 3(5). Retrieved from https://rjpn.org/jetnr/papers/JETNR2505034.pdf.
[16] https://github.com/Saed-Alewaity/Bigdata-project/tree/main