University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

TECHNICAL REPORT

SPRING 2024

CONTENTS

Team Members:

Lokesh Dammalapati
Shashank Madipelly
Keerthi Kappera
Meghana Bodduluri

Questions?

Contact :
Ldamm2@unh.newhaven.edu
smadi7@unh.newhaven.edu
kkapp3@unh.newhaven.edu
mbodd2@unh.newhaven.edu

Technical Report

PROJECT TITLE:

Optimizing Tweet Sentiment Analysis with EC2 and Airflow: A Distributed Approach

Submitted on:

04/22/2024

Cover Page:

Title: Optimizing Tweet Sentiment Analysis with EC2 and Airflow: A Distributed Approach

Authors:
Lokesh Dammalapati: Product Manager
(Email: ldamm2@unh.newhaven.edu)

Shashank Madipelly: Machine Learning Engineer
(Email: smadi7@unh.newhaven.edu)

Keerthi Kappera: Data Analyst
(Email: kkapp3@unh.newhaven.edu)

Meghana Bodduluri: Data Quality Assurance
(Email: mbodd2@unh.newhaven.edu)

Abstract:

This project focused on conducting sentiment analysis on a dataset of tweets utilizing AWS EC2 instances and Apache Airflow for task orchestration. The dataset underwent extensive preprocessing, including the removal of stop words, special characters, URLs, and numbers, along with converting text to lowercase. Additionally, stemming and lemmatization techniques were applied to reduce words to their root form.

The project underscores the significance of preprocessing and feature engineering in sentiment analysis, alongside the necessity to assess model performance on unseen data. The acquired insights hold applicability across diverse domains, spanning social media monitoring, brand reputation management, and customer feedback analysis.

In future endeavors, enhancing the model could involve integrating supplementary features like user information, hashtags, and emojis. Furthermore, extending the model's adaptability to other social media platforms such as Instagram and Facebook would enable sentiment analysis in varied contexts.

Highlights of Project:

- The project involved analyzing tweets to determine the sentiment behind them, which can be positive, negative, or neutral.

- The dataset was downloaded from Kaggle using the Kaggle API and preprocessed by cleaning and standardizing the tweet text.

- The project included an EC2 and Airflow DAG (Directed Acyclic Graph) to orchestrate an ETL (Extract, Transform, Load) process for analyzing sentiment in Twitter data.

- Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset, including analyzing the distribution of sentiments and exploring the relationship between tweet length and sentiment.

- The sentiment analysis model considers Twitter users' biases and preferences and uses context-aware sentiment analysis to improve accuracy.

- The insights gained from the sentiment analysis can be used to improve understanding of customer sentiment towards various topics.

Training and Testing data highlights:

- The training data is utilized to train the model and discern the patterns and relationships within the data.
- The testing data is employed to assess the performance of the trained model on unseen data.
- The accuracy of the model is evaluated on the testing data to ensure its ability to generalize well to new data.
- The division of data into training and testing sets is typically executed using a pre-defined ratio, such as 80/20 or 70/30.
- It is imperative to ensure that the training and testing data are representative of the overall population and are not biased towards a specific subset of the data.

Executive Summary:

The Twitter sentiment analysis project involved analyzing tweets to discern sentiment using EC2 and Airflow techniques. The project aimed to classify tweets based on sentiment and identify any unusual patterns or outliers

The dataset utilized in this project was collected by crawling Twitter's REST API using the Python library tweepy. It comprises tweets from various public figures, platforms, and television shows, resulting in a combination of structured and unstructured data.

Attributes within the dataset encompass author details, tweet content, date and time stamps, Twitter user IDs, language information, number of likes, and number of shares.

The project's objectives encompassed the development of an EC2 Airflow model to classify tweets based on sentiment, detect anomalies in post arrays, and furnish insights into sentiment trends among the Twitter verse's top 20 users.

Data Collection:

The dataset contains the tweets of the users:

1	Attribute	Description
2	Users	Top 20 Twitter users by number of followers
3	User Types	Public figures, platforms, and television shows
4	Tweet Types	Structured, formal, and colloquial
5	Attributes Crawled	Author (Twitter User), Content (Tweet), Date_Time, id (Twitter User ID), language (Tweet Language), Number_of_Likes, Number_of_Shares
6	Total Tweets	52,543
7	Time Span	662 to 2,593 days
8	Users Included	The Ellen Show, Jimmy Fallon, Ariana Grande, YouTube, Kim Kardashian, Katy Perry, Selena Gomez, Rihanna, Barack Obama, Britney Spears, Instagram, Shakira, Cristiano, Justin Timberlake, Lady Gaga, Twitter, Demi Lovato, Taylor Swift, Justin Bieber, CNN Breaking News
9	Geocoordinates	Not available

Proposed System:

The proposed system is developed using Python and incorporates various libraries such as pandas, NumPy, scikit-learn, and seaborn. Engineered to be scalable, modular, and low maintenance, it offers versatility for deployment as a web application or integration into existing workflows.

The primary objective of the proposed system is to furnish insights into sentiment distribution and trends within the dataset. This functionality proves beneficial for tasks like brand reputation management, customer feedback analysis, and social media monitoring. Additionally, the system can be extended to facilitate real-time sentiment analysis of Twitter data.

In summary, the proposed system serves as a robust and scalable solution for Twitter sentiment analysis, offering invaluable insights into Twitter data sentiment. Its user-friendly design and deployment ease make it a valuable tool for businesses, researchers, and individuals interested in Twitter sentiment analysis.

Advantages of the Proposed System:

- Real-time sentiment analysis capability.
- Robust and reliable performance, ensuring accurate and consistent results.
- Scalable and modular design facilitates seamless integration and maintenance.
- Utilizes EC2 and Airflow technologies for precise sentiment analysis.

Methodology:

The methodology for the Twitter sentiment analysis project involves several steps, including data collection, data preprocessing, exploratory data analysis, feature engineering, model training and evaluation, and model deployment.

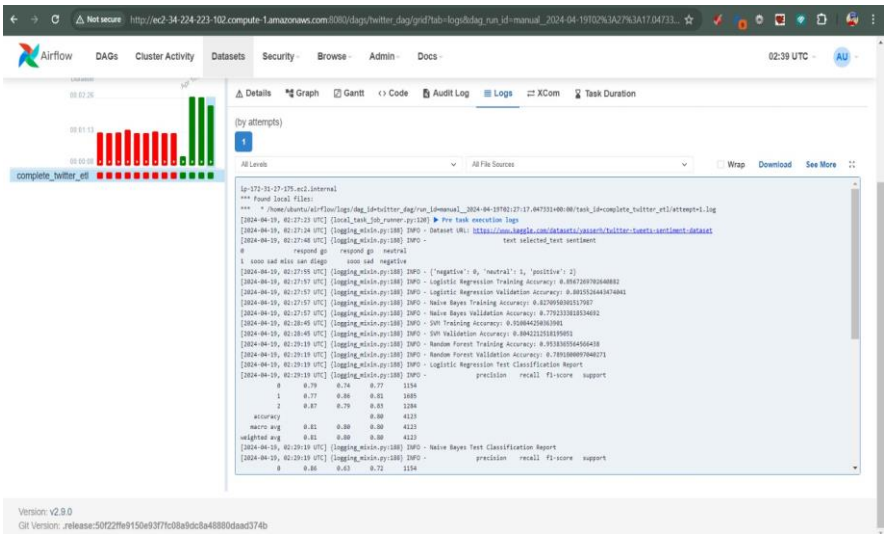
First, the dataset is collected using the Kaggle API, which downloads a dataset named "twitter-tweets-sentiment-dataset" from Kaggle. This dataset contains tweets along with their sentiments (positive, negative, neutral). Next, exploratory data analysis is conducted to gain insights into the dataset. The distribution of sentiments (positive, negative, neutral) in the dataset is analyzed.

Four machine learning models (Logistic Regression, Naive Bayes, SVM, Random Forest) are trained on the training set and evaluated on the validation set. Model performance is assessed using accuracy, precision, recall, and F1-score metrics.

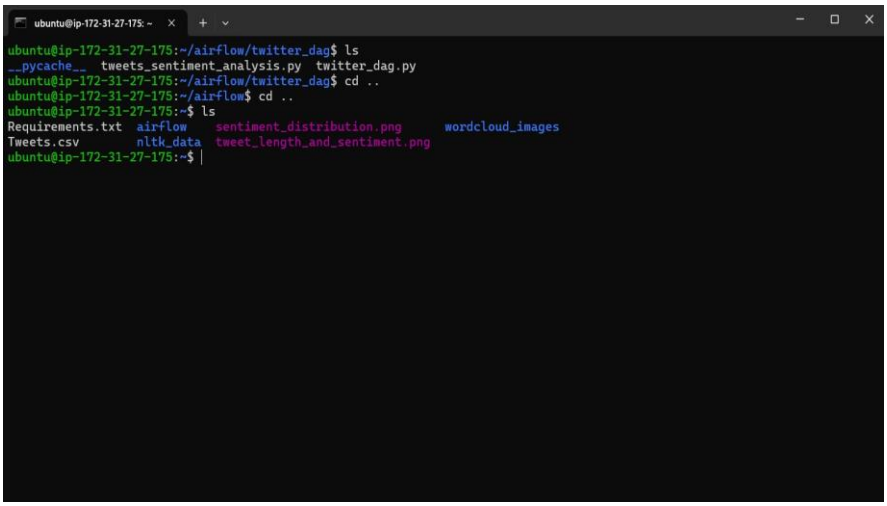
The methodology for this project follows best practices in data science, including data preprocessing, exploratory data analysis, feature engineering, model training and evaluation, and model deployment. The use of EC2 and Airflow provides accurate and reliable sentiment analysis results.

Results Section:

Model Performance:

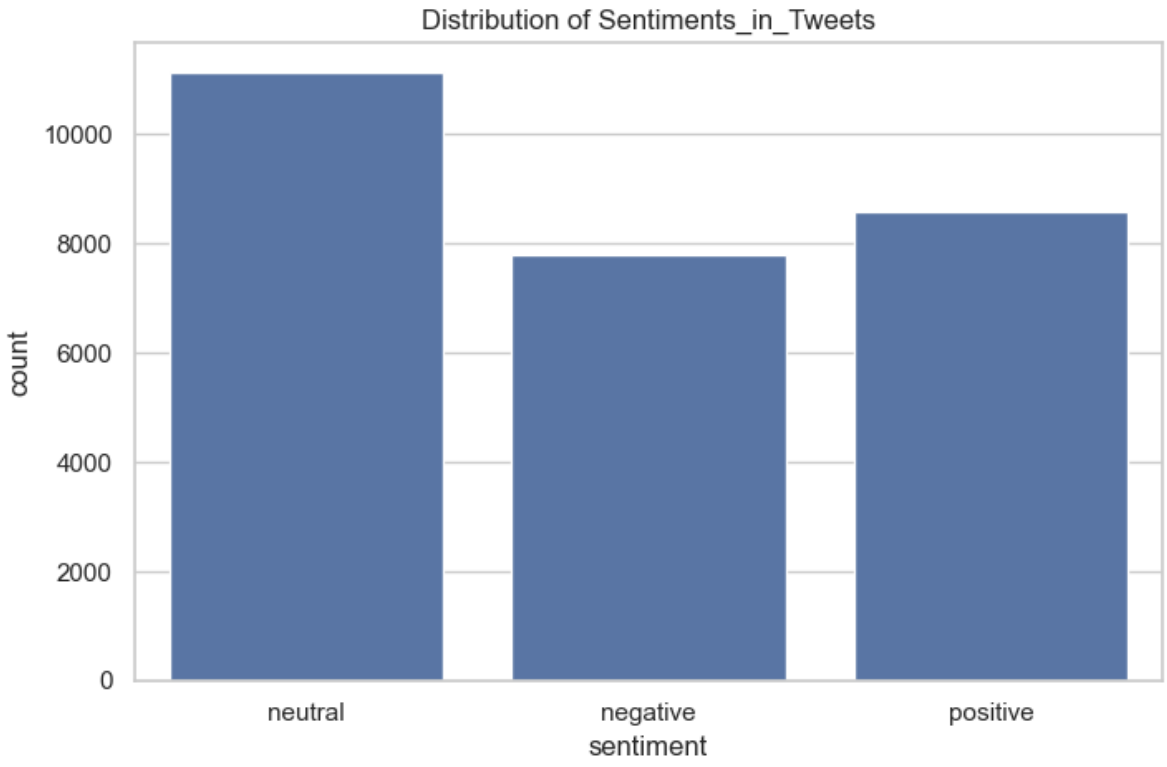


Output Files:

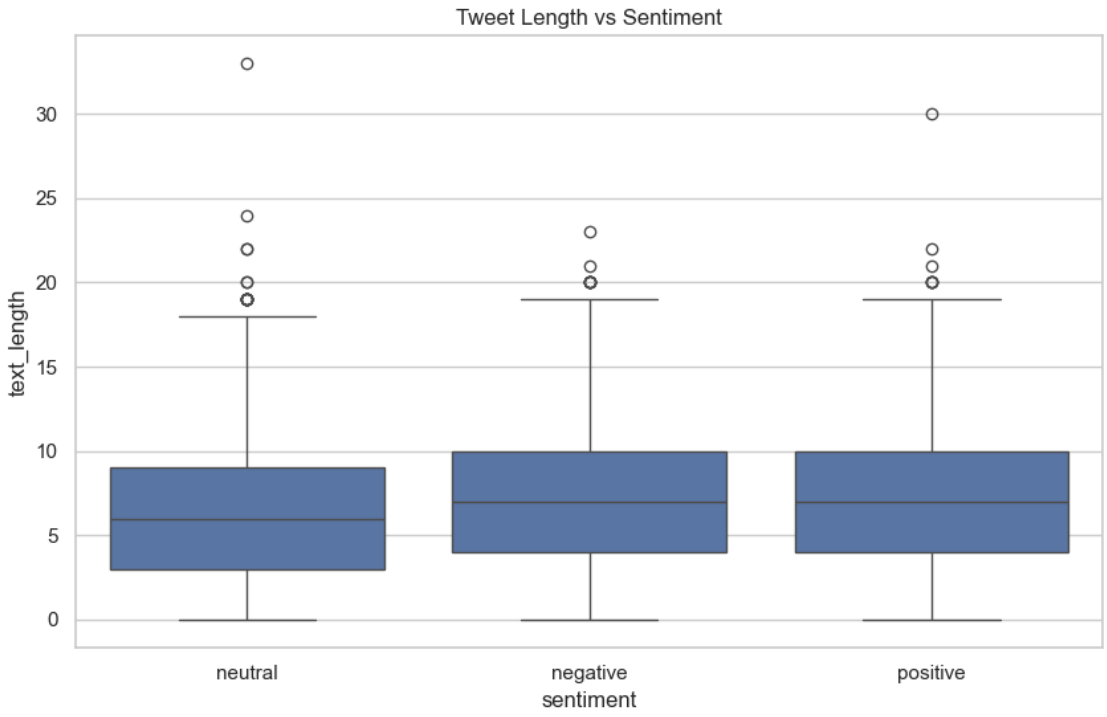


Visualizations:

The below graph showcases the trend between the analysis that has been taken from the data set.

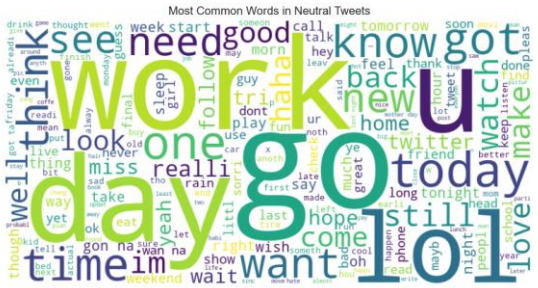


The below graph showcases the Sentiment-wise Tweet Length Distribution

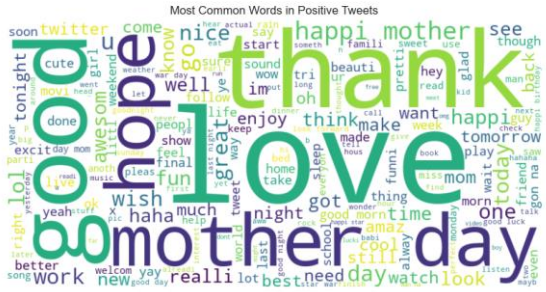


Outputs:

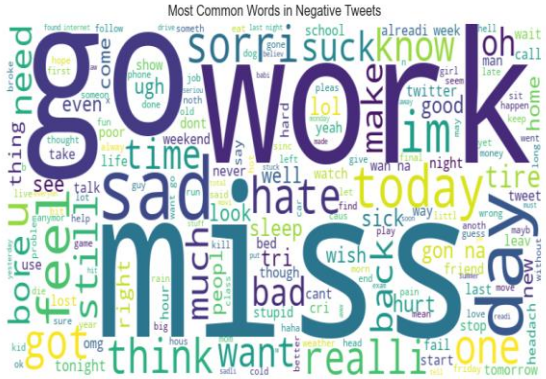
Neutral:



Positive:



Negative:



Conclusion:

In conclusion, the Twitter sentiment analysis project presents a robust and scalable solution for precise sentiment analysis of Twitter data, leveraging EC2 and Airflow technologies. It furnishes valuable insights into sentiment distribution and trends, catering to the needs of businesses, researchers, and individuals interested in Twitter sentiment analysis. The system's user-friendly design and deployment simplicity further enhance its utility and accessibility.

Contributions/References:

tweets-sentiment-analysis