# BUILDING AN NLP MODEL FOR ENTITY-LEVEL SENTIMENT ANALYSIS ON TWITTER

# DSF-PT08: FINAL PROJECT SUBMISSION

**Project Owner: James Wachira Muthee**

**Technical Mentor: Daniel Ekale**

# Table of Contents

# Business Problem Statement

Businesses struggle to extract meaningful insights from social media due to the lack of entity-specific sentiment analysis. Traditional sentiment analysis classifies entire messages without focusing on a given entity. This project aims to develop an entity-level sentiment analysis model to classify tweets as Positive, Negative, Neutral or Irrelevant concerning a specific entity, helping businesses track sentiment trends and improve decision-making.

# Business Objectives

The purpose of this project is to create an entity-level sentiment analysis model that can accurately classify tweets as Positive, Negative, Neutral or Irrelevant regarding a given entity. This model will help businesses and organizations to:

- Enhance Brand and Reputation Management
- Improve Customer Engagement Strategies
- Support Data-Driven Decision-Making

# Data

The Sentiment Analysis dataset used in this project was obtained from Kaggle

# Modelling Tools used

I used inbuilt libraries in Visual Studio Code tool to:

1. Load the datasets
2. Understand the datasets(shape, distribution and summary statistics)
3. Understand the data
4. Do exploratory data analysis
5. Do Data Preprocessing
6. Build Models
7. Evaluate Models' Performance
8. Validate the Model

# Data Understanding

Upon loading the data, the following observations were made:

- The training dataset had 74681 rows and 4 columns while the validation dataset had 999 rows and 4 columns
- The training dataset had 686 missing values in the tweet column while the validation dataset had no missing values
- Both datasets had one column with integer data type and 3 columns with categorical data type
- Both datasets had 4 unique values on the Sentiment Column
- The column names were renamed to ID, Entity, Sentiment and Tweet

# Data Preprocessing

Raw data was transformed into a structured format suitable for modeling. The following transformations were done to the data:

1. Converting from categorical to numerical
2. Reducing words to their root form
3. Removing words that do not add meaning to the text
4. Reducing texts to individual words
5. Removing hashtags, commas and other expressions
6. Converting Text to Lowercase

# Modelling

For this project, 2 models were developed, namely:

1. Logistic Regression Model(Baseline Model)

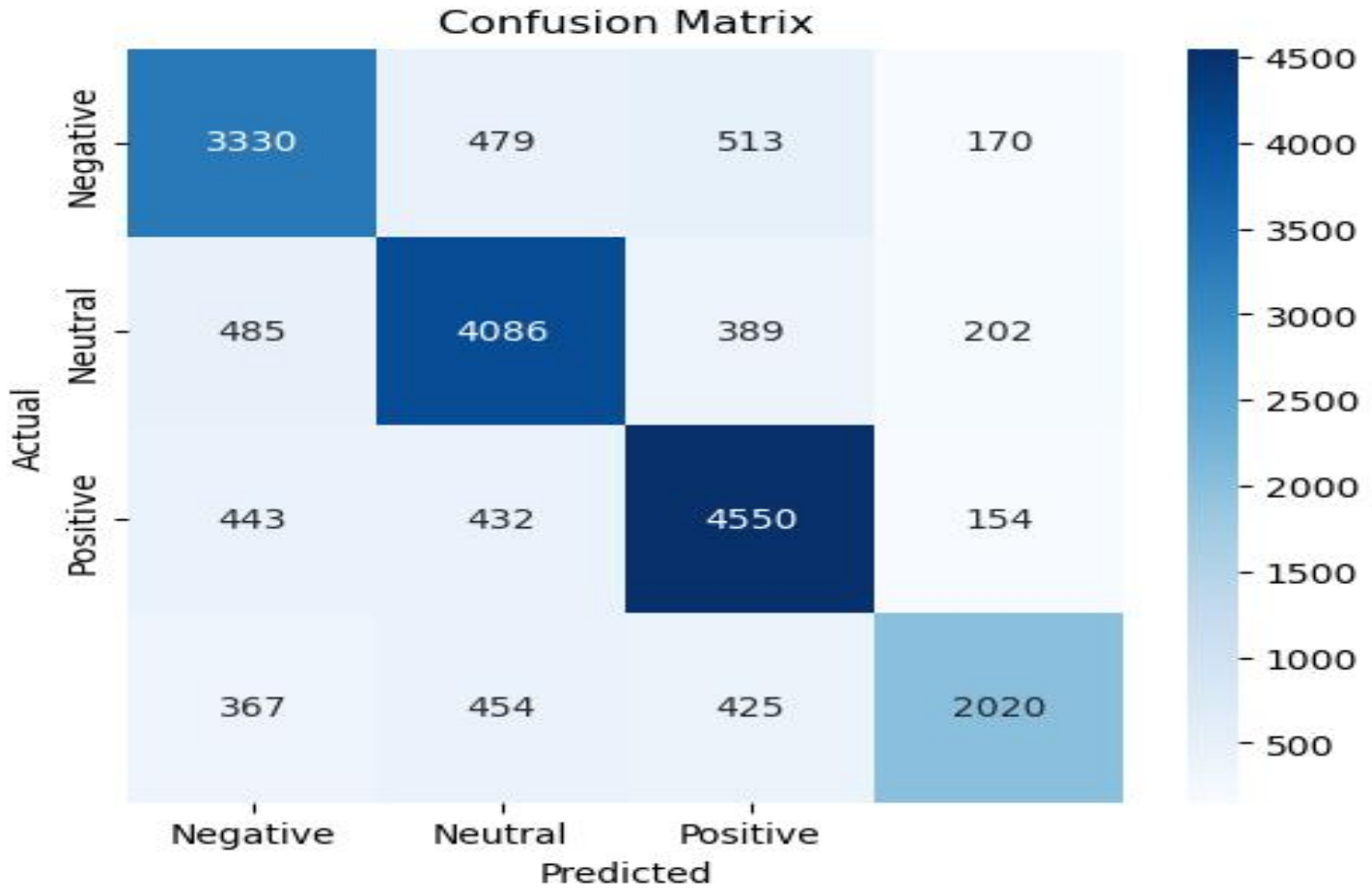
2. Random Forest Model

# Model Evaluation

The following Metric was used to evaluate the Models developed:

- **Accuracy Score**: It represents the proportion of correct predictions made by the model out of all predictions made.
- Model with highest accuracy score was taken as the best model

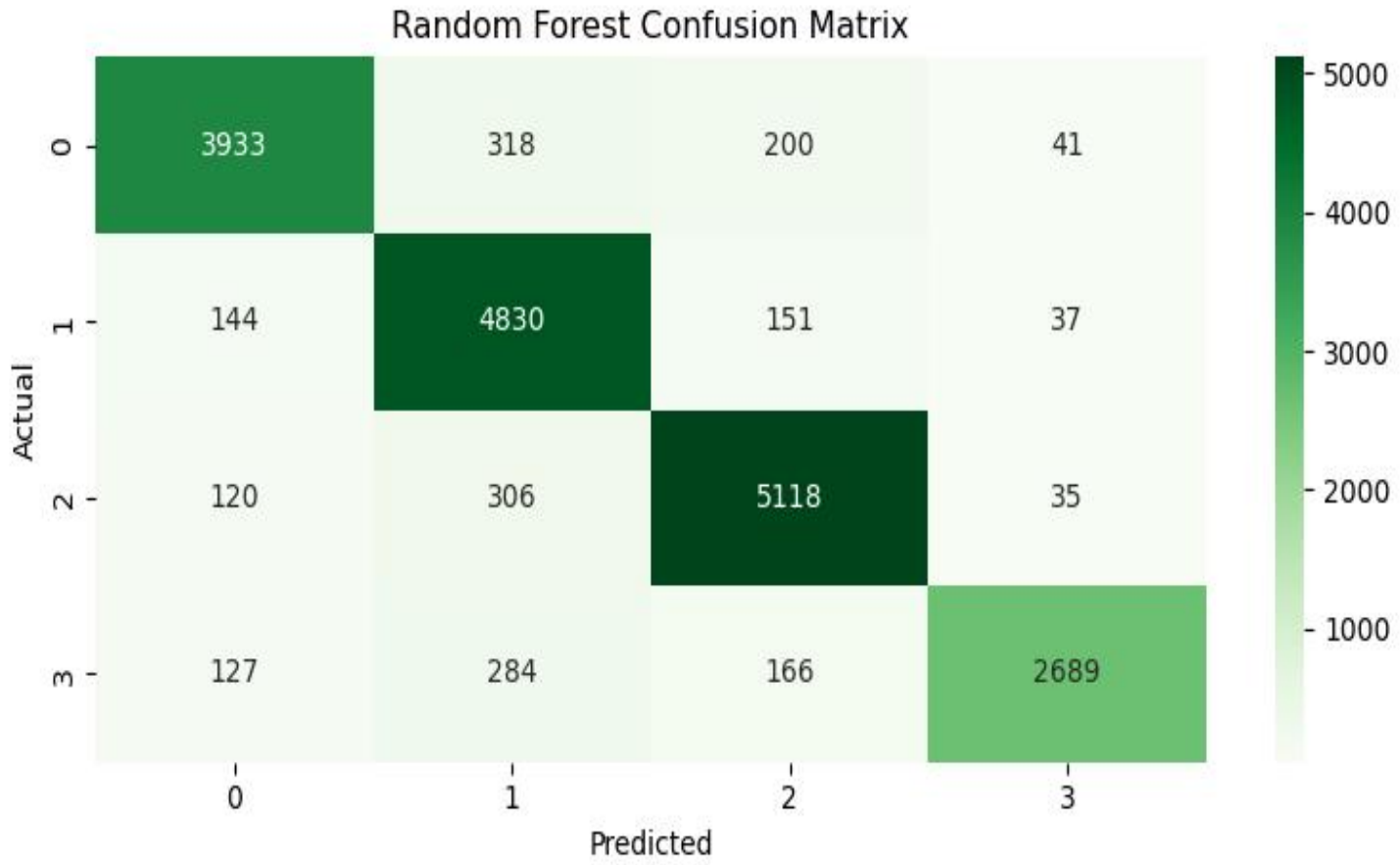
# Logistic Regression Model-Confusion Matrix




**Overall Metric Score**

Accuracy Score = 76%

# Random Forest Model-Confusion Matrix




**Overall Metric Score**

Accuracy Score = 89.6%

# Insights

1.**Logistic Regression Model**

- Model achieved an accuracy score of 76%
- The score indicates the model's effectiveness in sentiment classification.

**2. Random Forest Model**

- Model achieved an accuracy score of 89.6%. This shows its ability to capture complex sentiment patterns in tweets.
- It also suggests that the model is more adept at identifying sentiment nuances compared to logistic regression.

**3. Business Insights**

- Random Forest model can serve as a reliable tool for analyzing sentiment toward products, brands, and public figures.
- The ability to of the model to classify tweets as Positive, Negative or Neutral will enable businesses to refine their marketing strategies, manage their brand reputation, and enhance customer engagement.

# Conclusion

The overall objective of this project which was to develop an entity-level sentiment analysis model that can accurately classify tweets as Positive, Negative, Neutral or Irrelevant was achieved.

I concluded that the **Random Forest Model** with an **Accuracy Score of 89.6%** is the best model for this particular Sentiment Analysis Classification task.

# Thank You

# Q&A