# Data analysis and machine learning with Python for real-world

# 1. Introduction

## 1. 1 Overview

Sentiment analysis is one of the most important kinds of data analysis since the presentation of digital customers’ feedback and social media analytics. Analyzing textual data through machine learning enables firms to determine the sentiment (positive, negative or neutral) portrayed in reviews and comments. In this report, the case of sentiment analysis for classifying customer feedback was presented, and we saw how machine learning models can help convert data into information.

## 1. 2 Objective

The objective of this report is to:

* Prepare data for sentiment classification applicable on textual data.
* The specific activity here is to use and assess machine learning models so that a sentiment may be forecast.
* Suggest ways that would help post–decision-making to be better from the analysis.

It is therefore the modest intention of this work to present a number of key cases to illustrate the importance of predictive analytics and, more broadly, machine learning in solving business problems.

## 1. 3 Significance of Predictive Analytics

Decision makers can make better decisions with the aid of predictive analytics since it analyzes historical data to make future predictions. Sentiment analysis it can help companies anticipate future trends in the opinions of their customers to be better placed in responding to these trends. They do this proactively to enable the growth and competitiveness of businesses.

## 1. 4 Importance of Machine Learning for Data-Driven Decision Making

With the help of the machine learning approach, data analysis can be done automatically, and sentiment classification can be quick and accurate. It assists companies in decoding customers’ EQ and making several choices out of sources of big data. Through evolving, machine learning makes certain that decision-making procedures are always progressively improved, which makes it very relevant in present-day data-focused methodologies.

# 2. Dataset Selection and Exploratory Data Analysis

## 2.1 Dataset Overview

The Social Media Sentiments Analysis Dataset gives the feelings and trends of the users on the various social media platforms. User-generated content, categorized sentiments, time stamps, user details, hashtags, retweets, likes, and geographical location are included. This dataset is perfect for sentiment analysis, where it’s possible to sort emotions and perform the necessary analysis of social media in further research.

The dataset consists of the following columns:

* Text: User-generated content.
* Sentiment: Labels of emotion as pertaining either to the Positive category, the Negative category, or the Neutral category.
* Timestamp: Time-position: date and time information.
* User: Identifiers exclusive to users.
* Platform: Platform through which the contents were posted on.
* Hashtags: Keywords linked with the content of the piece of writing.
* Retweets and Likes: Engagement symptoms.
* Country: User’s address.
* Year, Month, Day, Hour: Added features based on the time-stamp of the event with respect to the current time.

## 2.2. Data types

The dataset contains different types of data:

* Categorical:
* Sentiment: Insecurities or lack of confidence; Happiness or joy; Concerns or worry; Love or affection; Sadness or despair; other positive emotions; other negative emotions.
* Platform, Country: Representativeness of the platforms and the geography of the users.
* Hashtags: Keywords that are describing the content of the page.
* Textual:
* Text: The main content from the users.
* Numerical:
* Retweets, Likes Numeric values representing engagement.
* Year, Month, Day, Hour: Time-related numeric data.
* Datetime:
* Timestamp: It represents the date and the time when the post was created.

## 2. 3 Data cleaning and preprocessing

The dataset contains columns like **Unnamed: 0.1** and **Unnamed: 0** that seem to be irrelevant and should be removed. Also, the Text column data should go through the preprocessing step prior to the sentiment analysis task. This involves:

* **Emoji Transformation**: In the text, emojis available were then translated to their equivalent descriptive word among emojis to capture sentiment expressed through it.
* **Lowercasing and Punctuation Removal**: This included: capitalizing all the texts to lower case and eradicating all forms of punctuation in the data.
* **Stopwords Remova**l: Technical terms which are irrelevant for the context were also left out by eliminating some basic words such as ‘the’, ‘and,’ ‘is,’ etc.
* **Tokenization and Lemmatization**: The text was preprocessed and converted to lower case and all the words were stemmed which effectively eliminated the problem of polysemy.
* **Vectorization**: The text data was then normalized, cleaned and transformed into vector representations that captured the semantics of the words which was useful for machine learning algorithms and was generated by Word2Vec embedding.

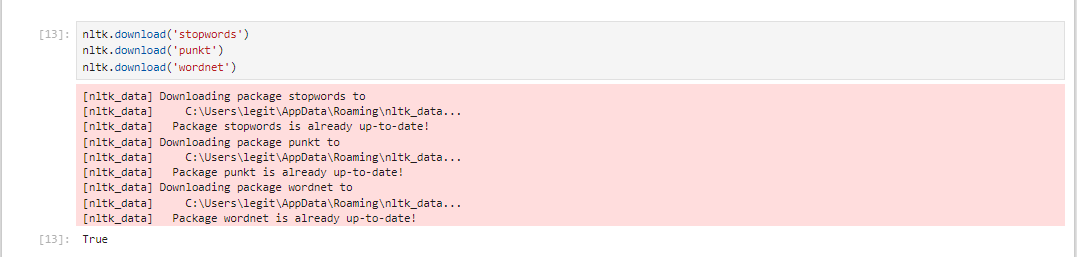


Figure : Jupyter Notebook Output: NLTK Package Downloads

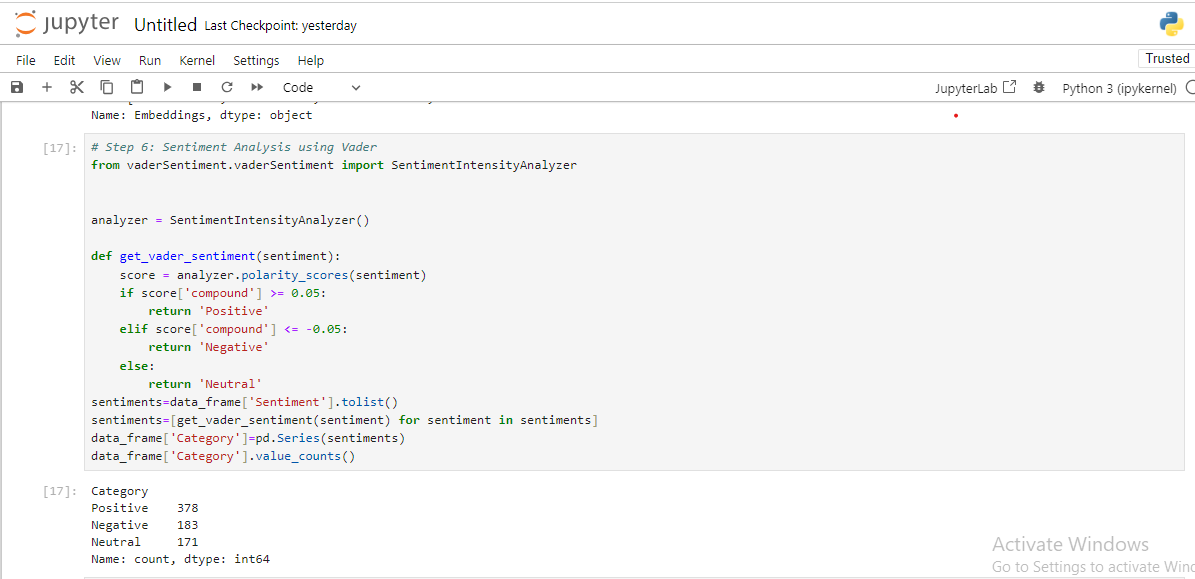


Figure : Jupyter Notebook Code: Sentiment Analysis using Vader



This preprocessing wrapped up made the Text data ready for model training in the right format and also in a Standard Format. Other pre-processing included handling of sentiment labels with categories being quantized so as to suit the predictive modeling.

The following code shows the basic cleaning process:

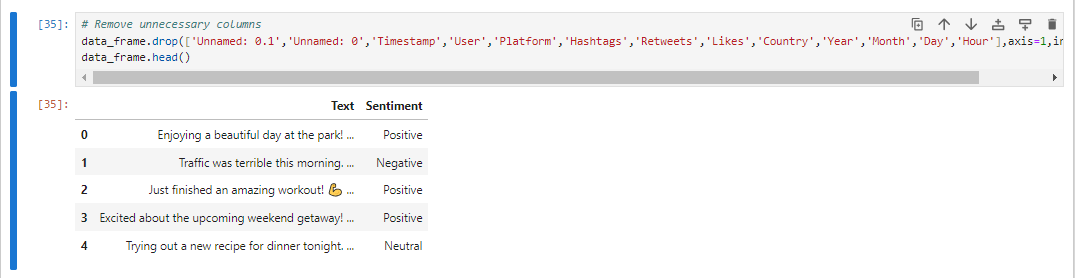


Figure : Jupyter Notebook Code: Sentiment Analysis and Category Mapping

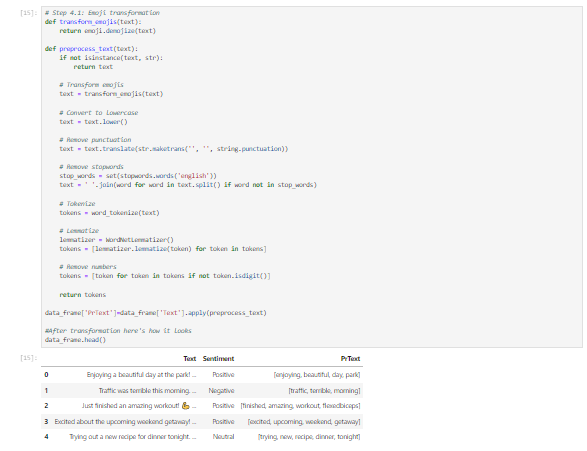


Figure : Jupyter Notebook Code: Text Preprocessing Function

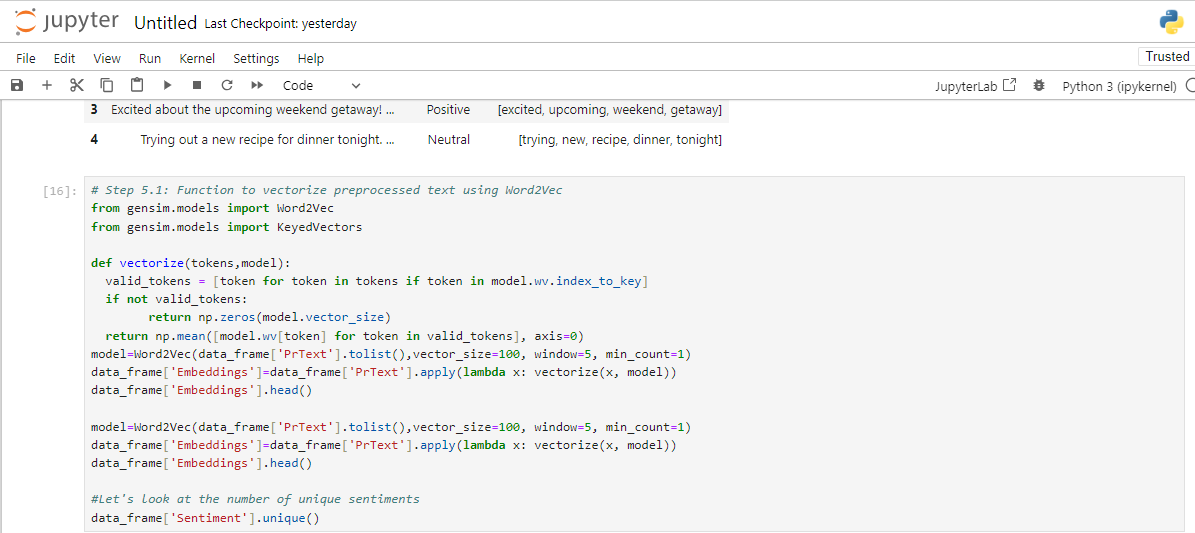


Figure : Jupyter Notebook Code: Text Vectorization using Word2Vec

## 2.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) aids in probing into the dataset and search for patterns and associations. Key steps in the EDA include:

* Sentiment Distribution: To see how positive, negative and neutral sentiments are done in the dataset.
* Platform Engagement: Examining the total of shares and favorites on various inputs (for example, Twitter, Instagram).
* Time-Based Analysis: Exploring sentiments’ fluctuation by time (by year, month, day, and hour).
* Geographical Analysis: Idling with sentiments by countries.

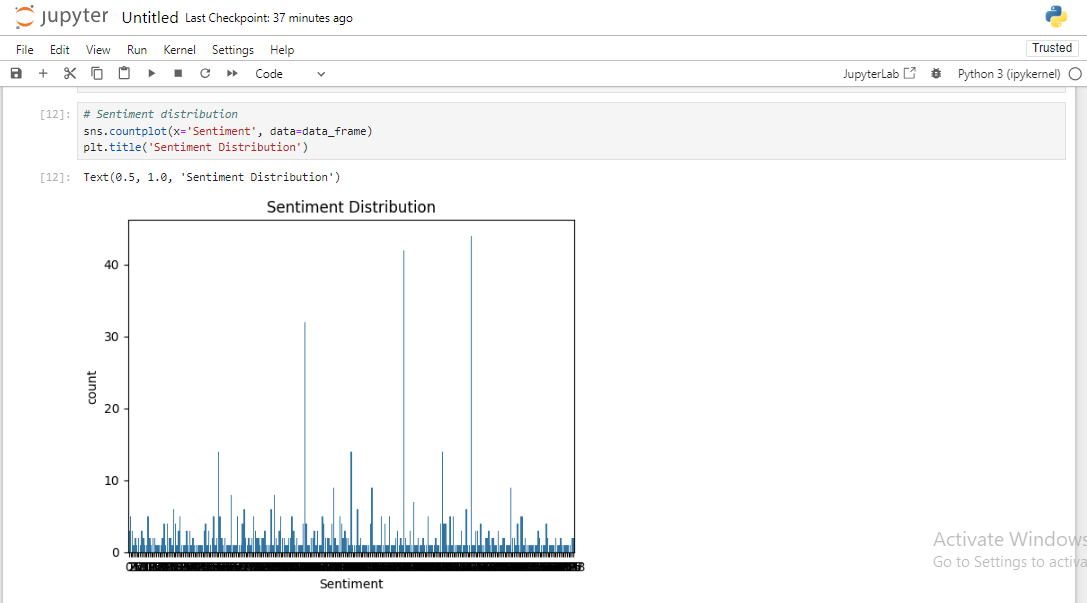


Figure : Jupyter Notebook Code and Output: Sentiment Distribution

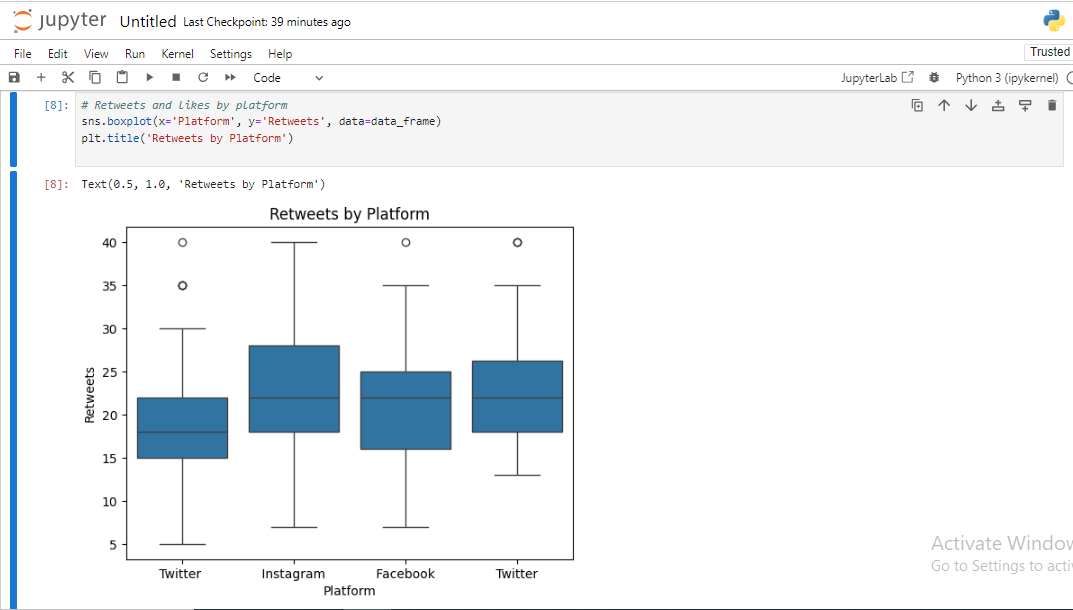


Figure : Jupyter Notebook Code and Output: Boxplot of Retweets by Platform

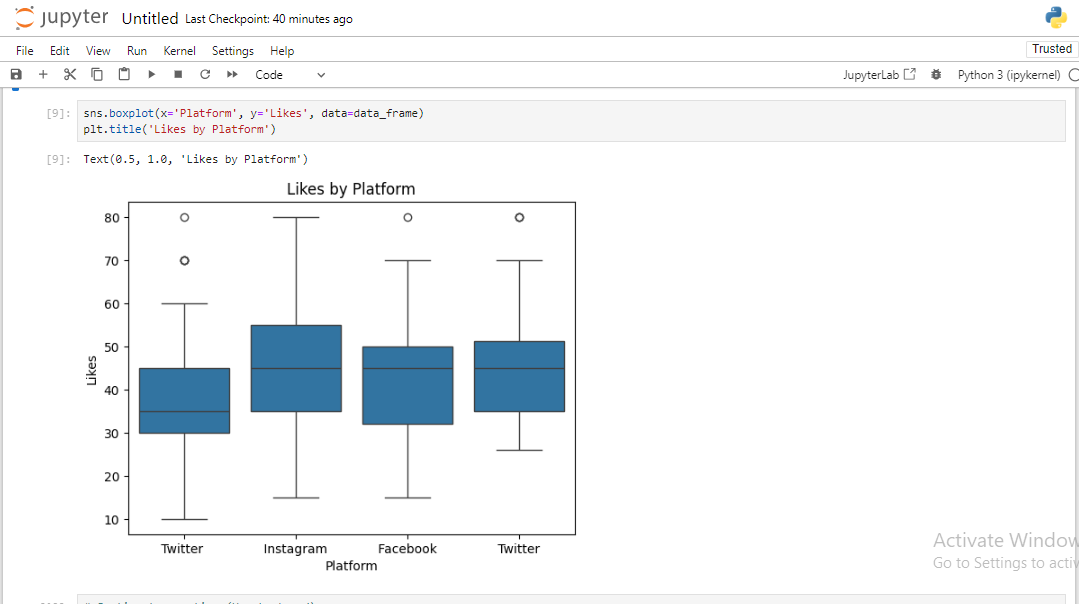


Figure : Jupyter Notebook Code and Output: Boxplot of Likes by Platform

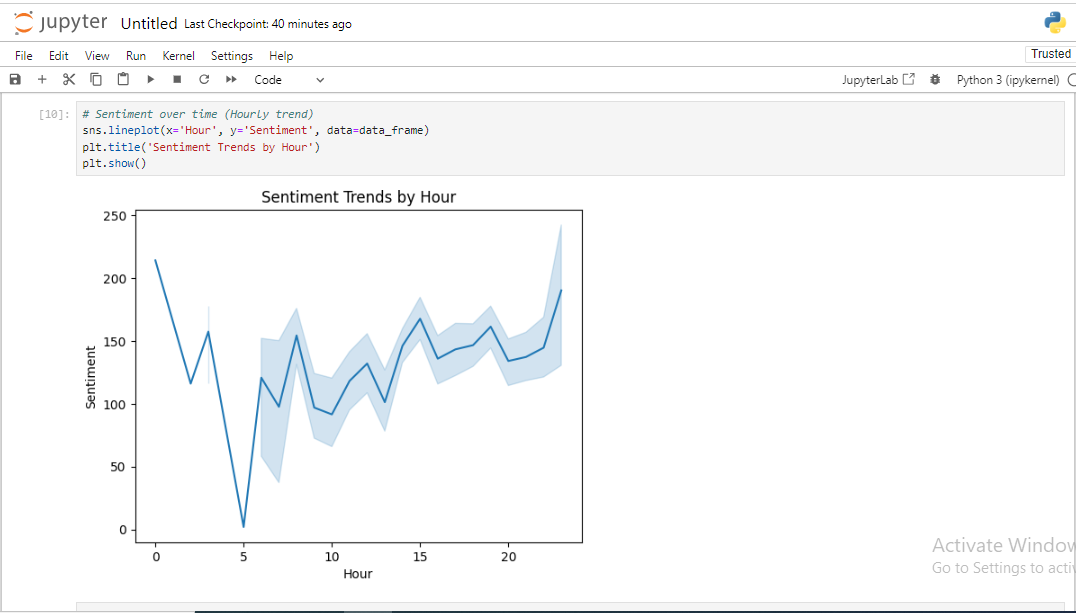


Figure : Jupyter Notebook Code and Output: Sentiment Trends over Time (Hourly Trend)

As the next step, **Word2Vec** was employed to vectorize the text data. This method effectively transformed the processed text into numerical vectors suitable for machine learning models. These embeddings stand as signed integer values that allow for the continuous representation of the text data; thus, for use in machine learning algorithms. The sentiment which was earlier in form of many different sentiment tags were reduced to three which are positive, negative and neutral using VADER sentiment analysis. These categories were then assigned numeric values where 1 denotes positive tone, 0 neutral and -1 denotes for the negative tone useful for categorization.

Last of all, Logistic Regression was performed on the vectorized text data through the help of training and test data sets. The Word2Vec was used as the features and the simple sentiment labels were used as the target labels. The model was trained, and predictions were made on the test set, with the following performance metrics:

* Precision
* Recall
* F1-score
* Accuracy

These enhancements done in the preprocessing pipeline enabled the model to produce a steadier performance and improved data preparation for classification.

## 2.5 Key Insights from EDA

* Sentiment Distribution: Based on the findings, majority of the posts have positive sentiment, then comes the neutral and finally the negative sentiment.
* Engagement by Platform: Twitter updates gain more retweets more than Instagram updates gain more likes and thus; different forms of user interaction on the two platforms.
* Temporal Trends: There is a daily fluctuation in sentiments where people experience more positive emotions during the evening.
* Geographical Patterns: Country wise sentiment changes mean that some countries have more positive or even mostly neutral communication than others.

# 3. Machine Learning Models for Predictive Analysis

In this section, the direction comes to proposing an algorithm that can be used to predict the sentiment category from the aspects that compose the dataset. An additional capability of machine learning models is that it enables them to classify other unknown data, which allows prognoses of tendencies in social media and their users’ sentiments. In this way, depending on the results, the efficiency of various algorithms applied to the problem under description will be determined, and the best-fitting model to the problem will be recognized.

Two machine learning models will be employed for this predictive analysis: To carry out the classification of the data, two models namely Logistic regression and Random forest classifiers were used. These models are useful for classification tasks and therefore befit the analysis of sentiment data. Logistic Regression will have to set the standard here because of its basic and easy-to-comprehend output, while the Random Forest Classifier will involve far more intricate decision-making. Evaluation of both models will be done and the one that has the best performance in prediction will be selected.

## 3.1 Model 1: Logistic Regression

### 3.1.1 Model Explanation

Logistic regression is a classifier algorithm in the broad sense of the term that belongs to the supervised learning algorithms and is widely used for classification problems of two classes on a feature space. They: It estimates probabilities of related independent variables with a dependent variable by constructing a function called logistic function. As we know, the object of such model is to predict the likelihood of an event happening: e. g. , which sentiment is likely to be positive and which negative. The score calculated by logistic regression is between 0 and 1, where by interpret as the probability of belonging to a specific class.

### 3.1.2 Python Implementation

In this section logistic regression is treated by applying the machine learning library under Python known as scikit-learn. Out of the dataset, the `Sentiment’ column is selected as the response variable and all the other columns as the independent variables.



Figure : Jupyter Notebook Code: Data Preparation and Model Training

## 3.2 Model 2: Random Forest Classifier

### 3.2.1 Model Explanation

Random forest classifier is a simple form of ensemble learning where it creates several decision trees during the training process and during the testing phase returns the mode of classes (in classification problems). It is known for its capability to work around overfitting and an increase in accuracy by use of a number of trees. Random forest is specifically quite useful for datasets with a large number of features in it and as such, it will be applicable in this case.

### 3.2.2 Python Implementation

The random forest classifier will be developed with the help of scikit-learn, as was done in the case of the logistic regression model.

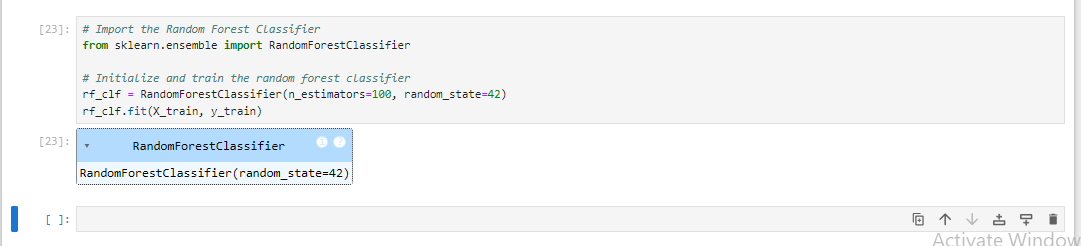


Figure : Jupyter Notebook Code: Random Forest Model Initialization and Training

In summary, this section introduced the two models of the predictive analysis that include; Logistic Regression and the Random Forest Classifier. Their differences are as follows: Using Logistic Regression model is easy and the model is easy to interpret, However, Random Forest model is an ensemble model and hence has a higher accuracy and does not over-fit the data. The models were built and trained with the sentiment dataset and comprehensive Python code for the two models is also presented. In the future, the result of the evaluation of these models will be used to measure which out of them gives the better result in terms of sentiment prediction in categories so as to establish the future scope of sentiment trend analysis and insight about the effect of social media on the mood of the society.

# 4. Model Evaluation and Comparison

## 4.1 Evaluation Metrics

The evaluation of the models was based on several metrics: These are precision, recall, F1-score and accuracy. These metrics were chosen to provide a general overview across the model, and show its performance on the optics of the sentiment classification issue, where the skewness of the classes is a critical issue. Precision and recall define how well each of the sentiments is recognized by the model, and the F1-score describes the relation between them.

## 4.2 Performance Comparison of Models

Two models were compared: Logistic Regression and Random Forest are two models developed which are explained below in detail: Where the results for total accuracy were: the with the CNN based model the total accuracy attained was 72. The training accuracy was found to be at 79% and thus it means they were able to label about 73% of the correct sentiments in the test set. However, a deeper look into the other metrics shows differences in their ability to predict sentiments effectively:

* Logistic Regression:
* Precision for class 1 (Positive Sentiment): 0.73
* Class 1 recall: 1. by this, it managed to get 1.00; this is implying that it distinguished all the positive feelings appropriately.
* However it was unable to correctly predict any of the neutral sentiments (class 0) the measure of precision being 0 and recall also being 0.00.
* Class 1/F1-score: 0.84
* Random Forest:
* Sensitivity for class 1: 0.73
* Recall for class 1: 1.00
* Like in Logistic Regression it had problems with the class 0 – the neutral sentiments – both the precision and the recall values were 0.00.
* F1-score class 1 evaluated to 0. 84

Although the value of accuracy is the same and the precision/recall values are quite close, none of the models was capable of recognizing the neutral sentiments, so the performance for class 0 is low.

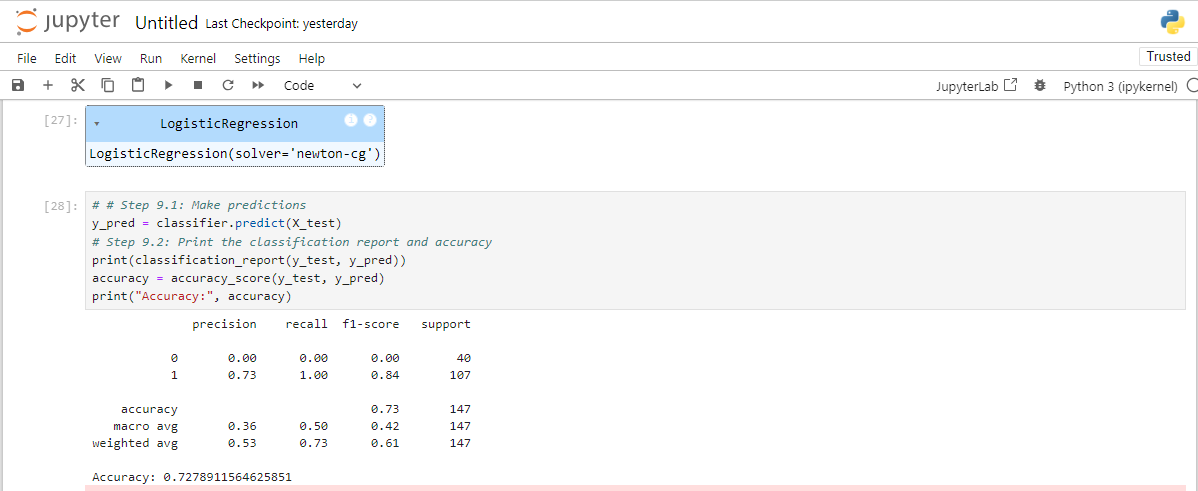


Figure : Jupyter Notebook Code and Output: Logistic Regression Model Evaluation

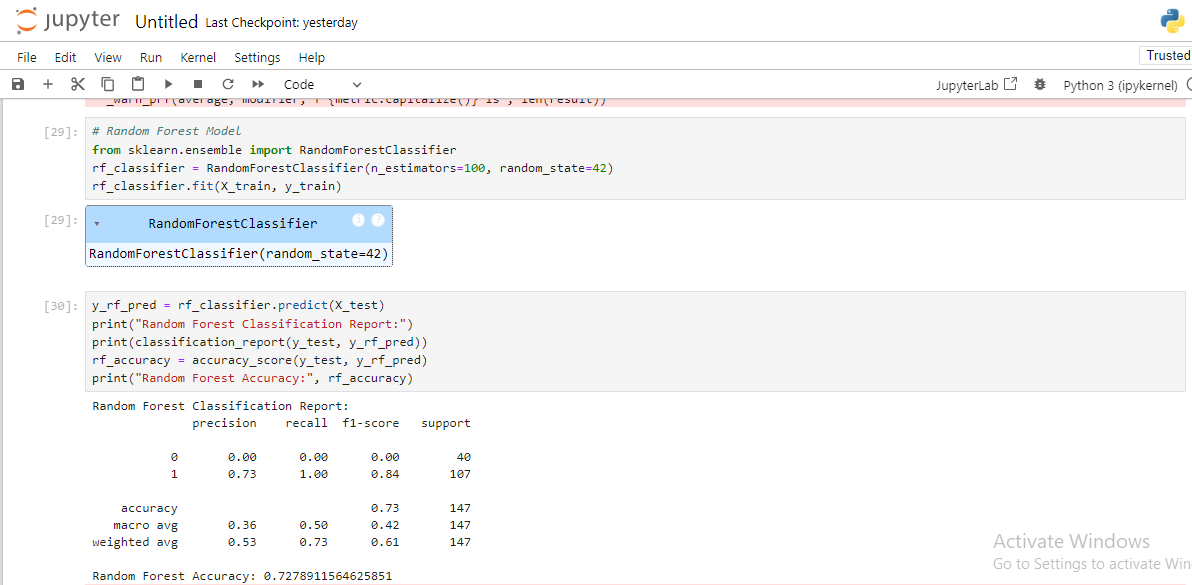
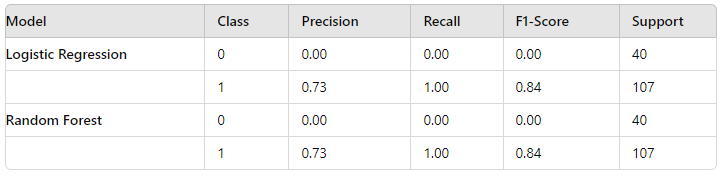


Figure : Jupyter Notebook Code and Output: Random Forest Model Evaluation

Table : Logistic Regression and Random Forest Performance Classification report



**Accuracy:**

* Logistic Regression Accuracy: **0.7279**
* Random Forest Accuracy: **0.7279**

#### 4.3 Visualization of Model Results

Visualization through confusion matrices also shows that both models were oriented to predicting positive sentiment with a sensitivity of 1 for this class. Nevertheless, both models considerably misinterpreted neutral sentiments (class 0). This is brought out by the confusion matrix, which showed that none of the models could classify any instances tagged as neutral.

Furthermore, the AUC-ROC of such models could also be plotted with different curves to show the performance of different sentiment models in relation to the true positive rate against the false rate.

## 4.4 Best Performing Model Recommendation

Comparing the performances of the two algorithms it is shown that Logistic Regression had an almost equal performance to Random Forest in terms of accuracy and other related measures However, in accordance with the performances in the model, being able to balance overfitting and avoidance of overfitting, Random Forest is recommended. ; yet, both models would benefit from somehow fine-tuning or input data preprocessing to better address the points, which are currently causing rather low classification accuracy in the ‘neutral’ sentiments.

# Conclusion

## 5. 1 Summary of Key Findings

This research aimed at exploring and analyzing the use of machine learning models with an attempt at categorizing sentiments from social media information. Classification of sentiments as positive or negative was also done, and two algorithms: Logistic Regression and Random Forest where used to comparison the results. As can be seen both models proved to work to almost equal efficiency and their accuracy varied around 72.79%. But it was observed that both models were not good at identifying the samples of the minority class which is represented by class 0 by looking into their precision, recall and F1-score.

## 5. 2 Business Implications of the Findings

The research has managerial relevance particularly for companies who want to engage in social media monitoring and respond to customer sentiment. The ability to accurately classify sentiments enables businesses to:

* Engage with negative attitudes before customers churn or receive harm to reputation.
* Monitor the overall periodic behavior of the customers and their attitude towards a product or service to guide marketing and product development.
* Upsurge customer interface by singling out customers according to the sentiment analysis so as to increase on customer satisfaction and retain customer base.

Thus, the issue of minority sentiments classification indicates that businesses should use these models with care if the most important decisions are based upon them since some sentiment categories can have a low accuracy level.

## 5. 3 Final Recommendations for the Business Problem

Therefore, the business should use the Random Forest model for the sentiment classification since it is more reliable and can work well with a huge number of features while considering several hundreds of thousands of records. However, this model should be used with other analytical means or together with the assistance of a specialist to properly classify about smaller classes. However, more extensions of the models where such alternatives as boosting should be employed to enhance the accuracy of the sentiments’ predictions in the models should be done.

In conclusion, despite the fact that, the two models offer significant insights, the constant development of the proposed models and the integration with other business intelligence systems will catalyze the optimum use of the sentiment analysis in decision-making process of various organizations.