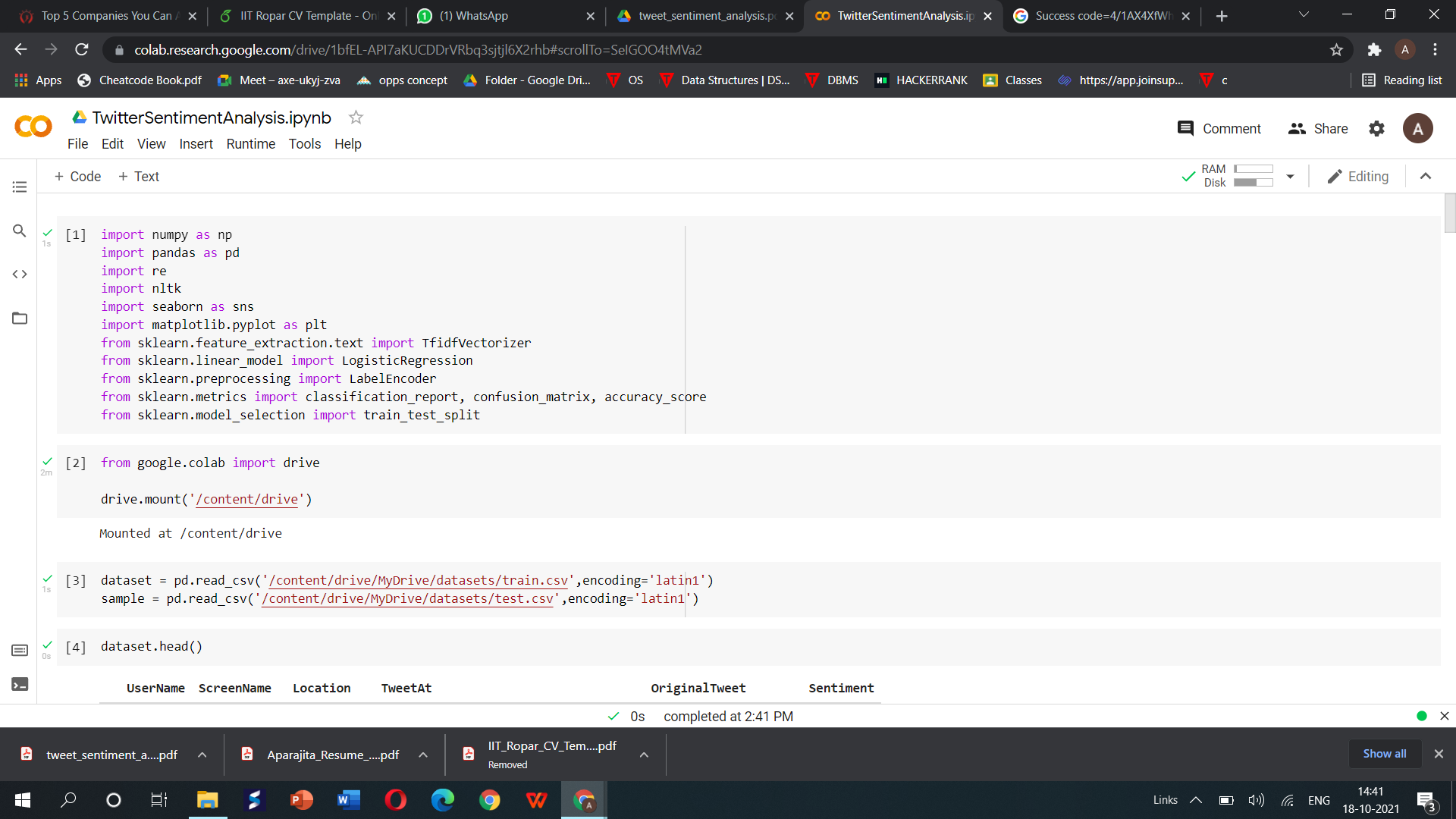
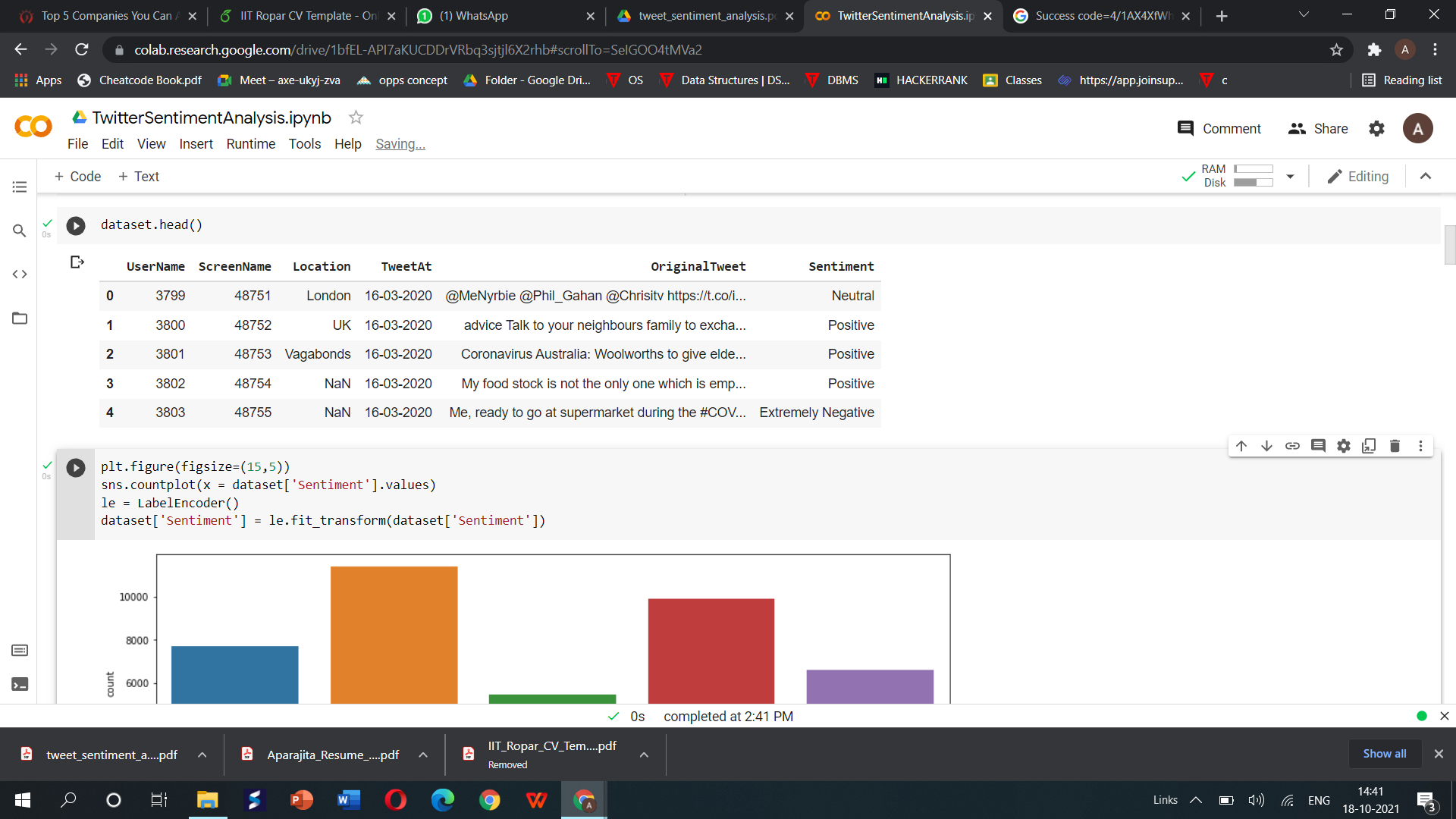
**SENTIMENT ANALYSIS IN TWITTER DATASETS**

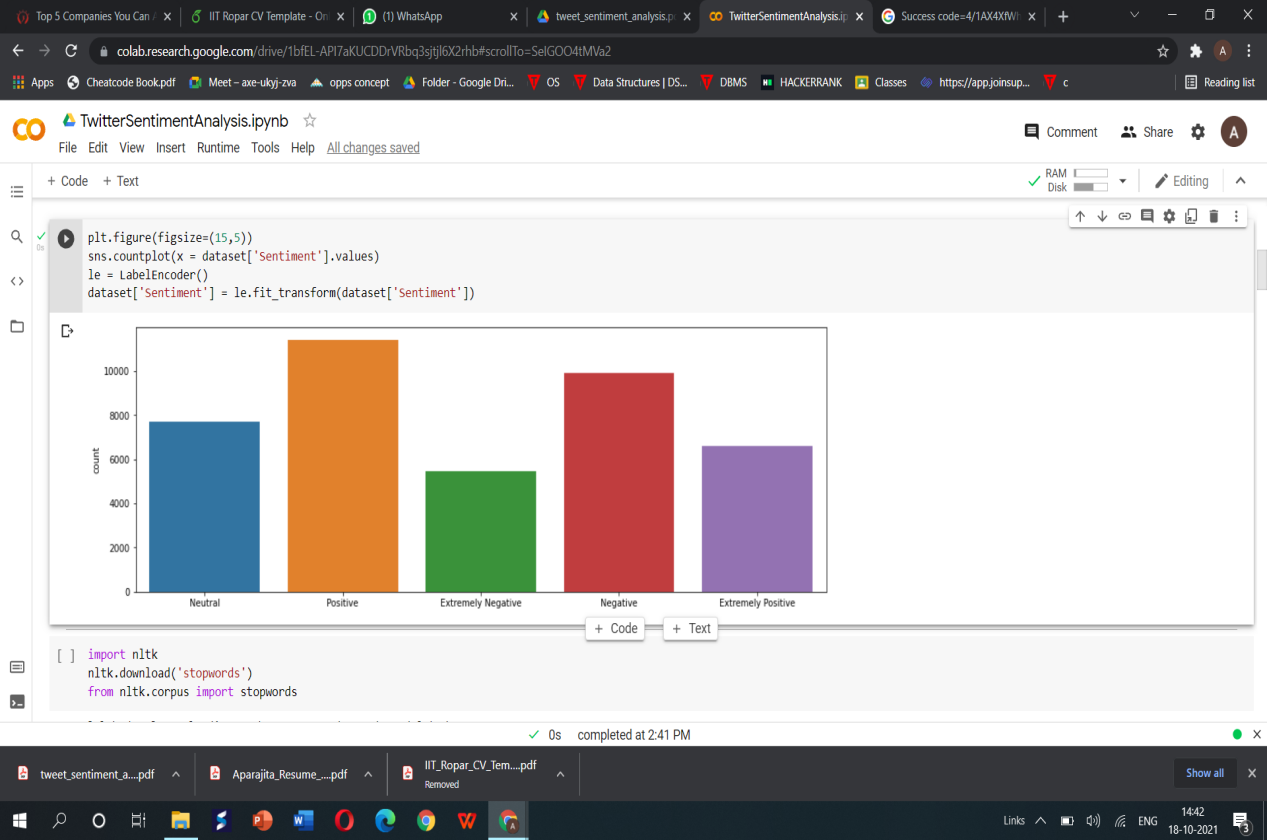
**OVERVIEW:**

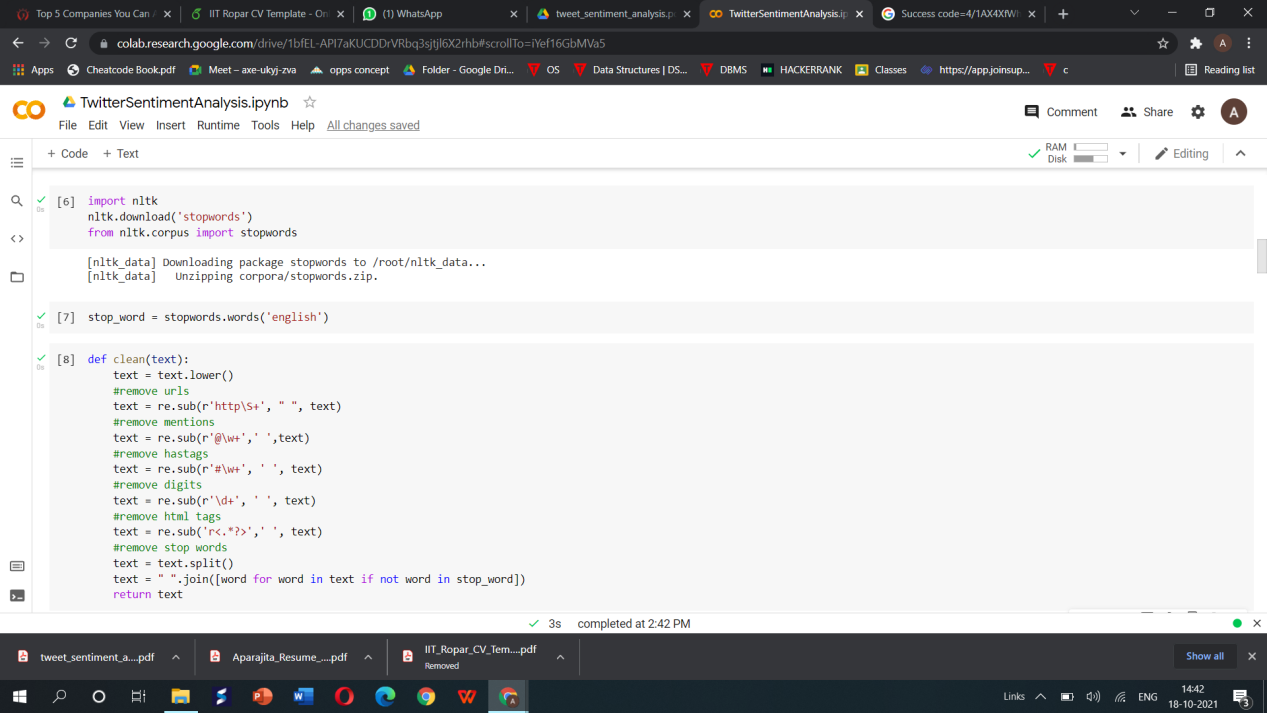
Sentiment Analysis is the automated process of determining whether a text expresses a positive, negative or neutral opinion about a product or topic. To analyse costumer sentiment, this process uses natural language processing, text analysis and statistics concepts. In this work, **Logical Regression** and **Random forest** models are used and their performances are compared on which Logical Regression gives more accuracy and performs better than Random Forest. To implement above models, Dataset is divided into training and testing dataset. Our work here is to process the training data consisting of tweets on various recent themes centred on current events and to predict the sentiment of the test data. There are a series of sentiments given in the training data. EDA approach (Exploratory data analysis) is used in this work to analyse and investigate datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what data can tell us before the modelling tasks. Implementation is done in **python3** using **Google Colab.**

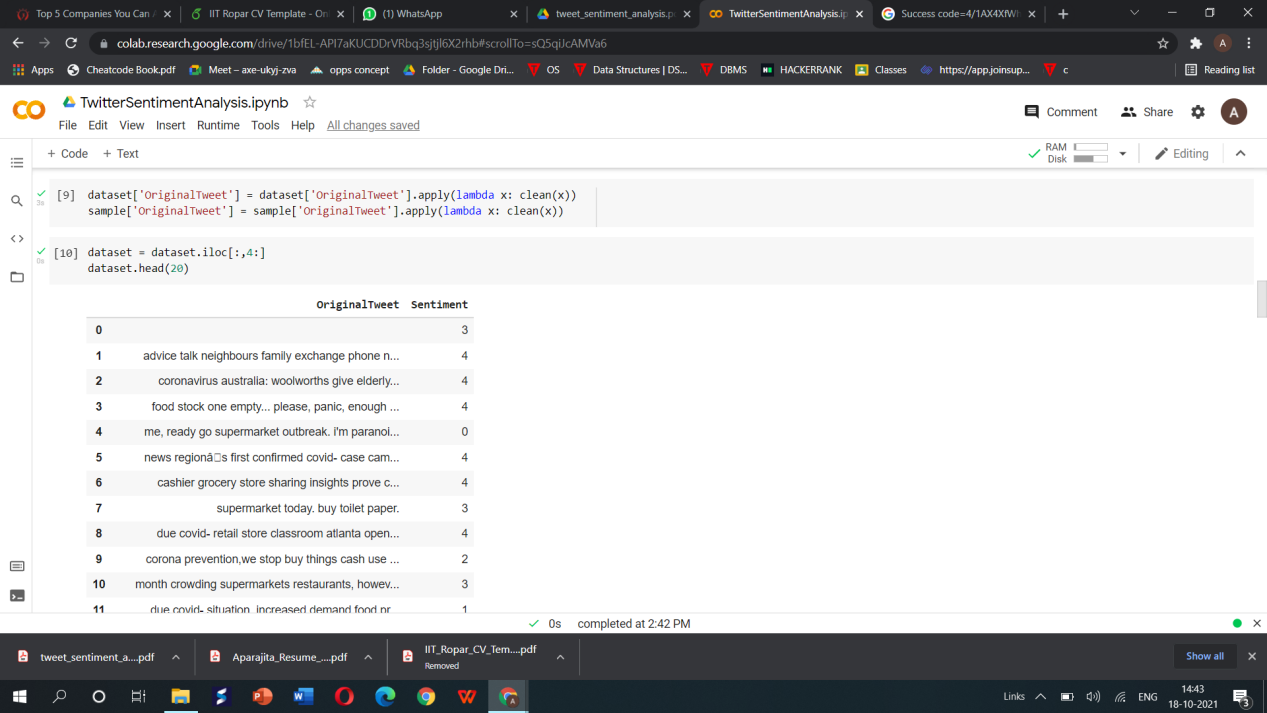
**SOURCE CODE AND OUTPUT:**

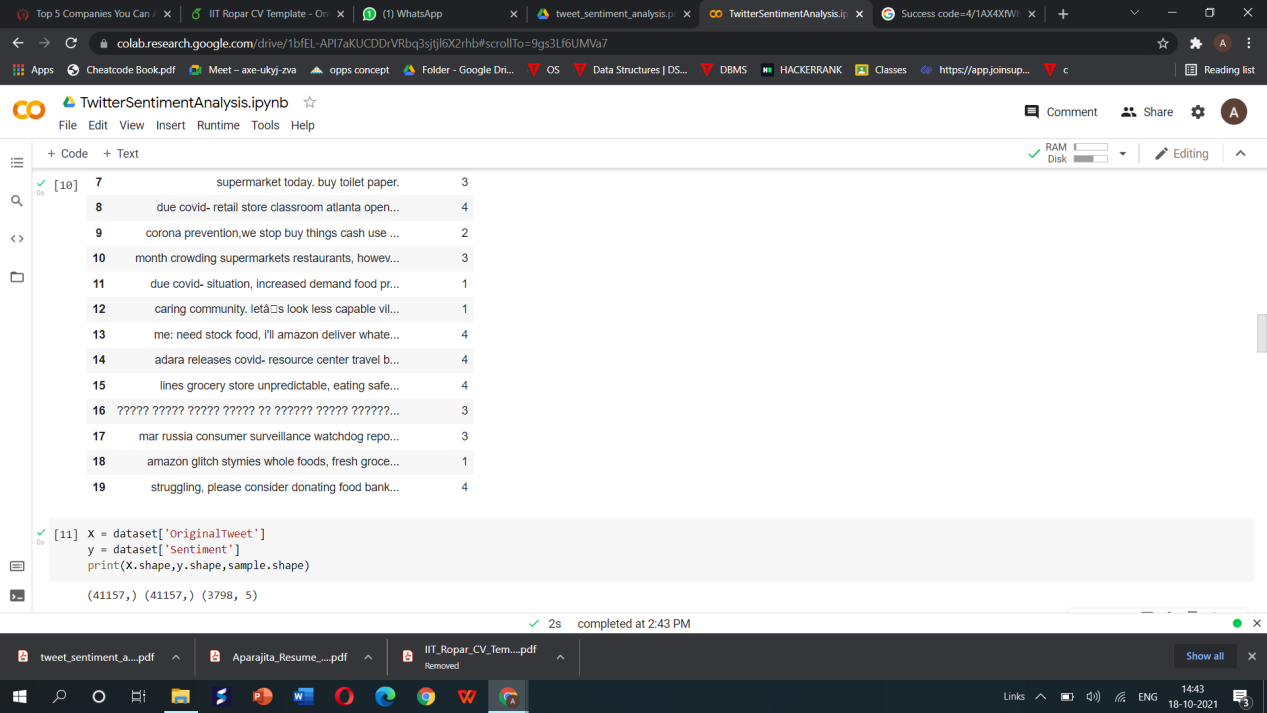
****

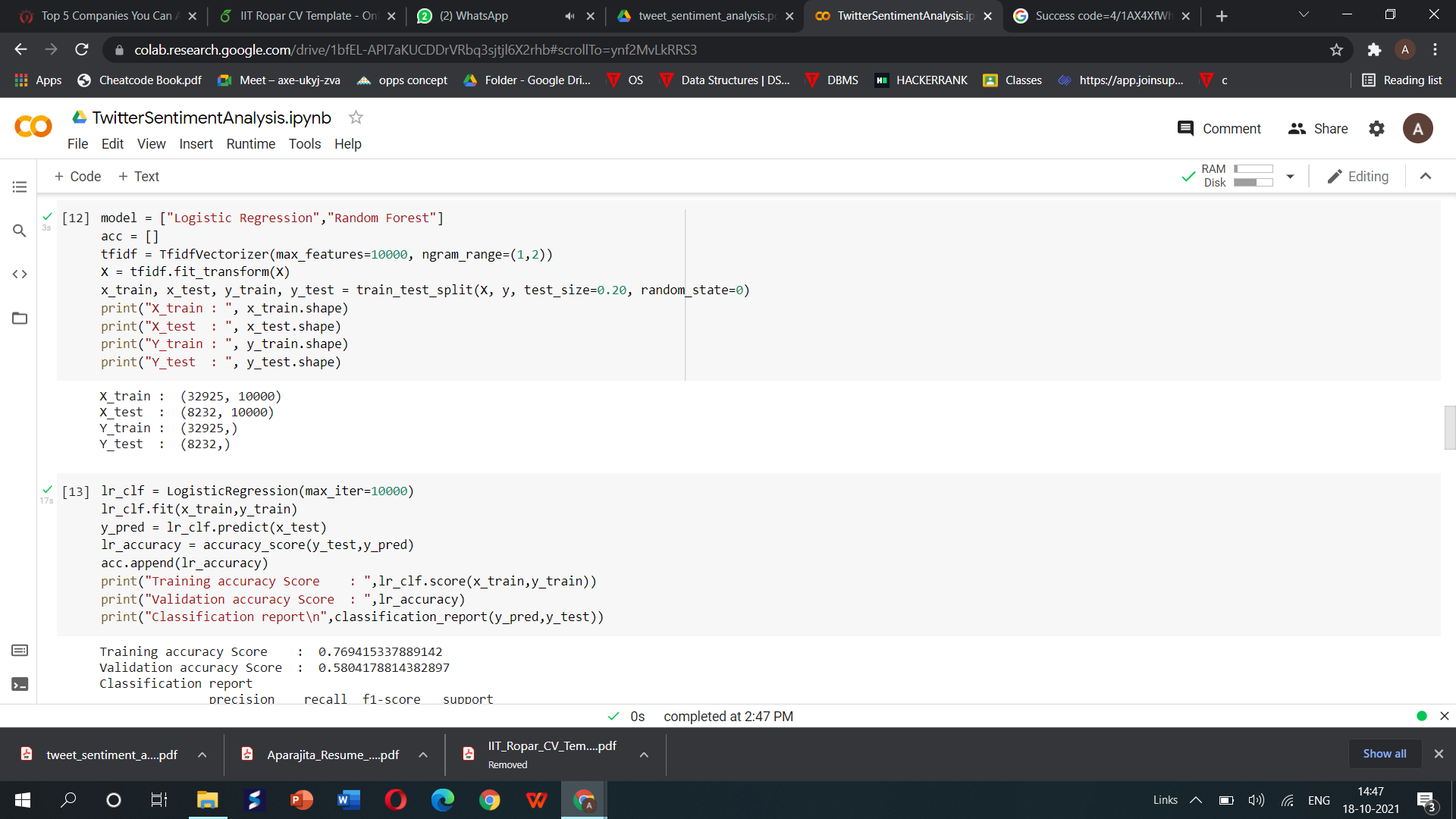


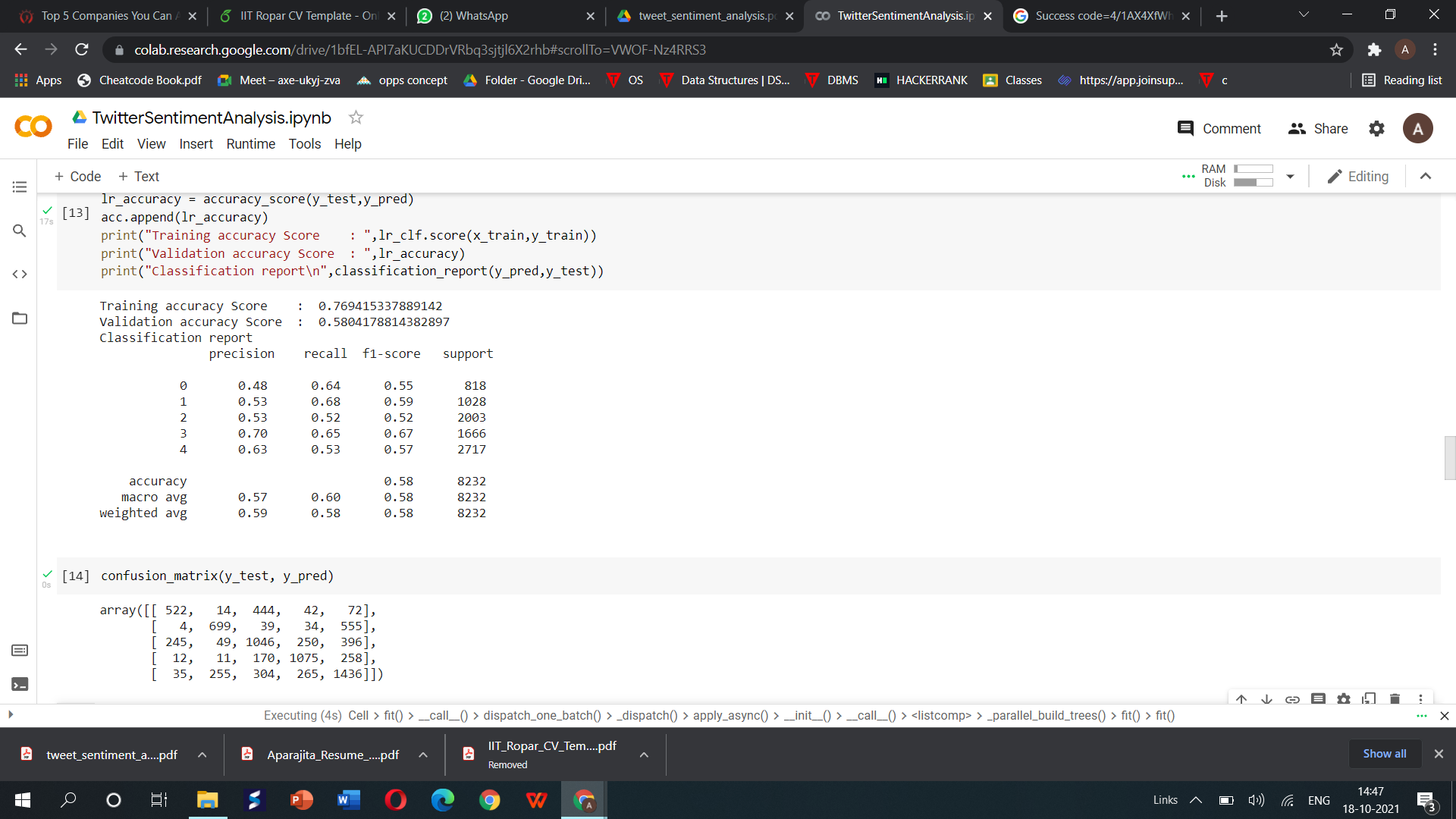


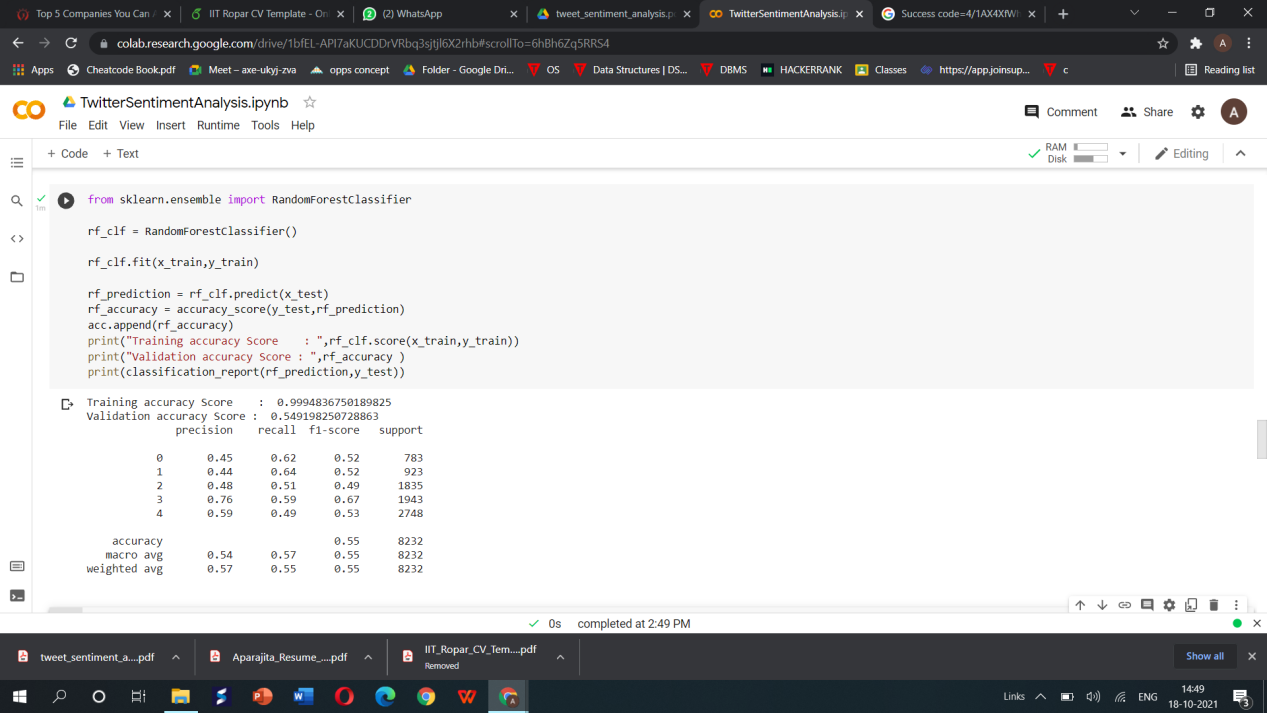
****

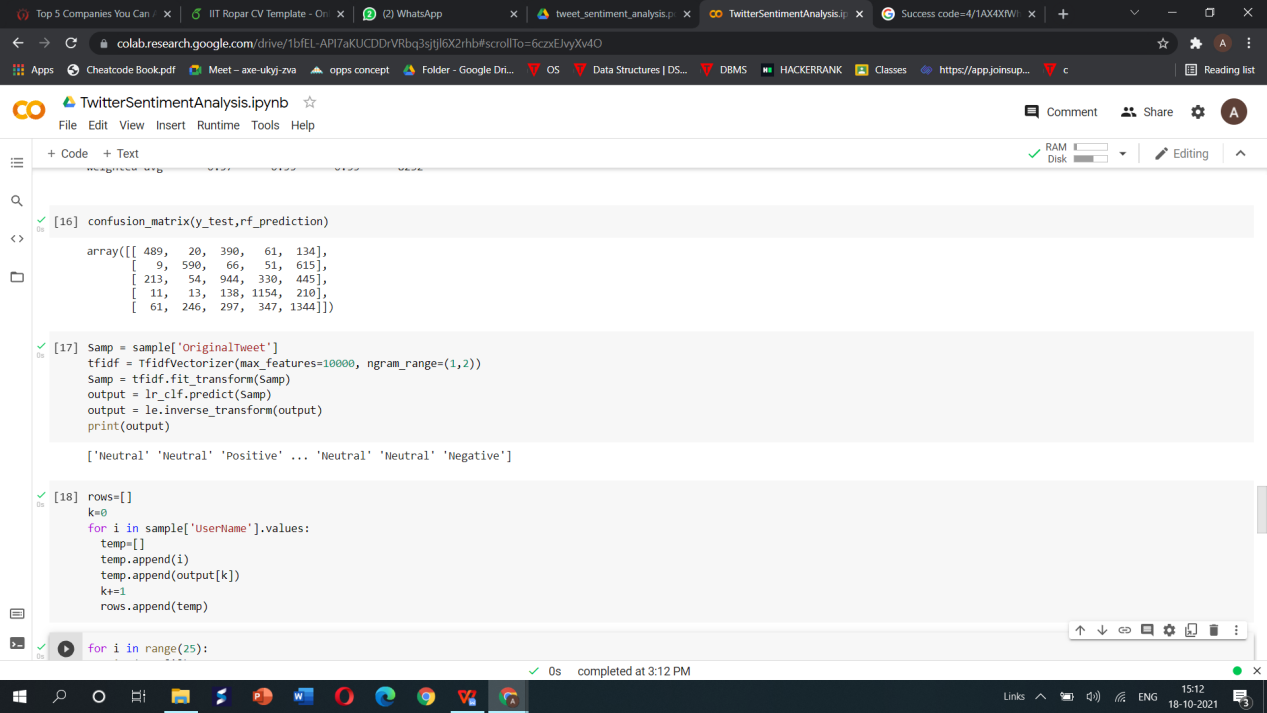
****

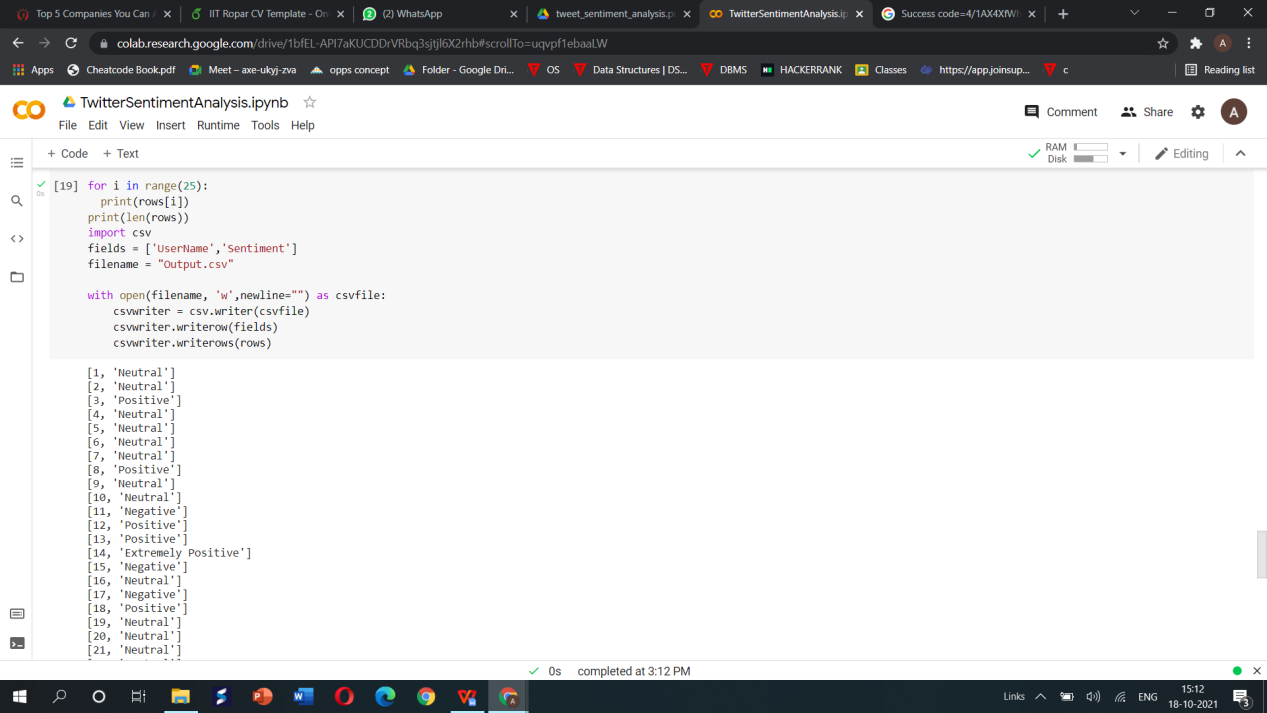
****

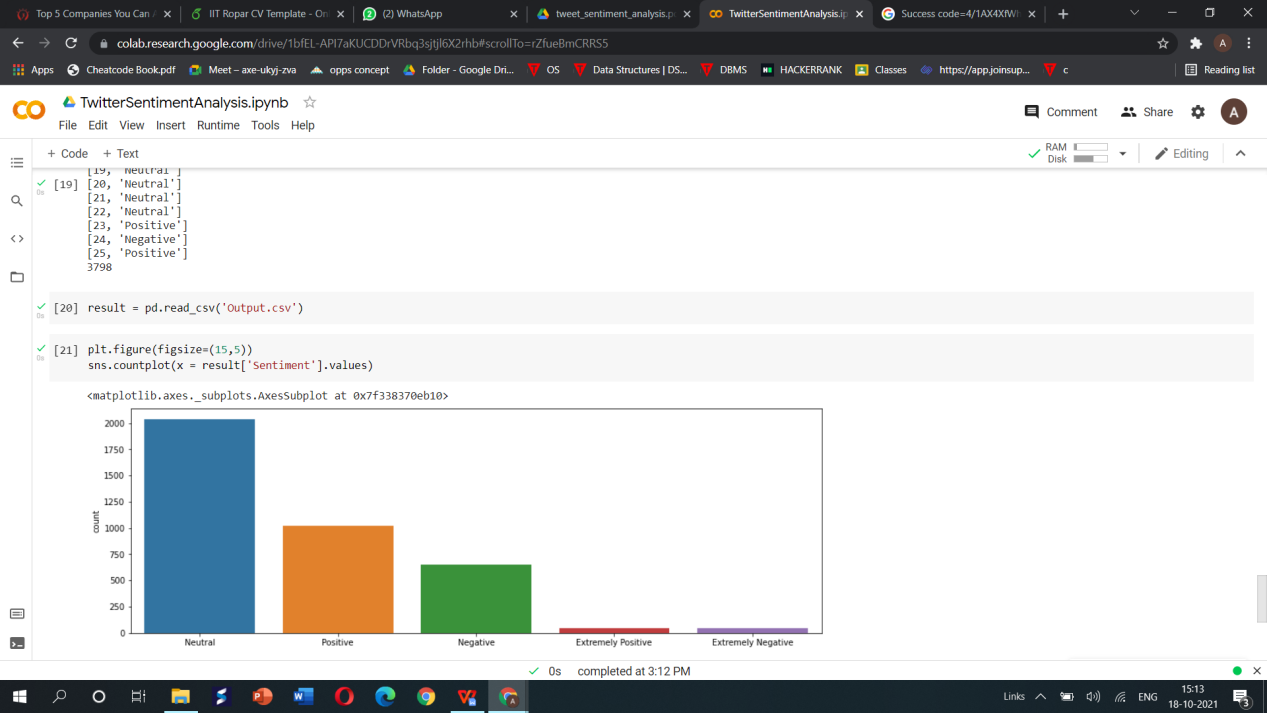
****

****

****

****

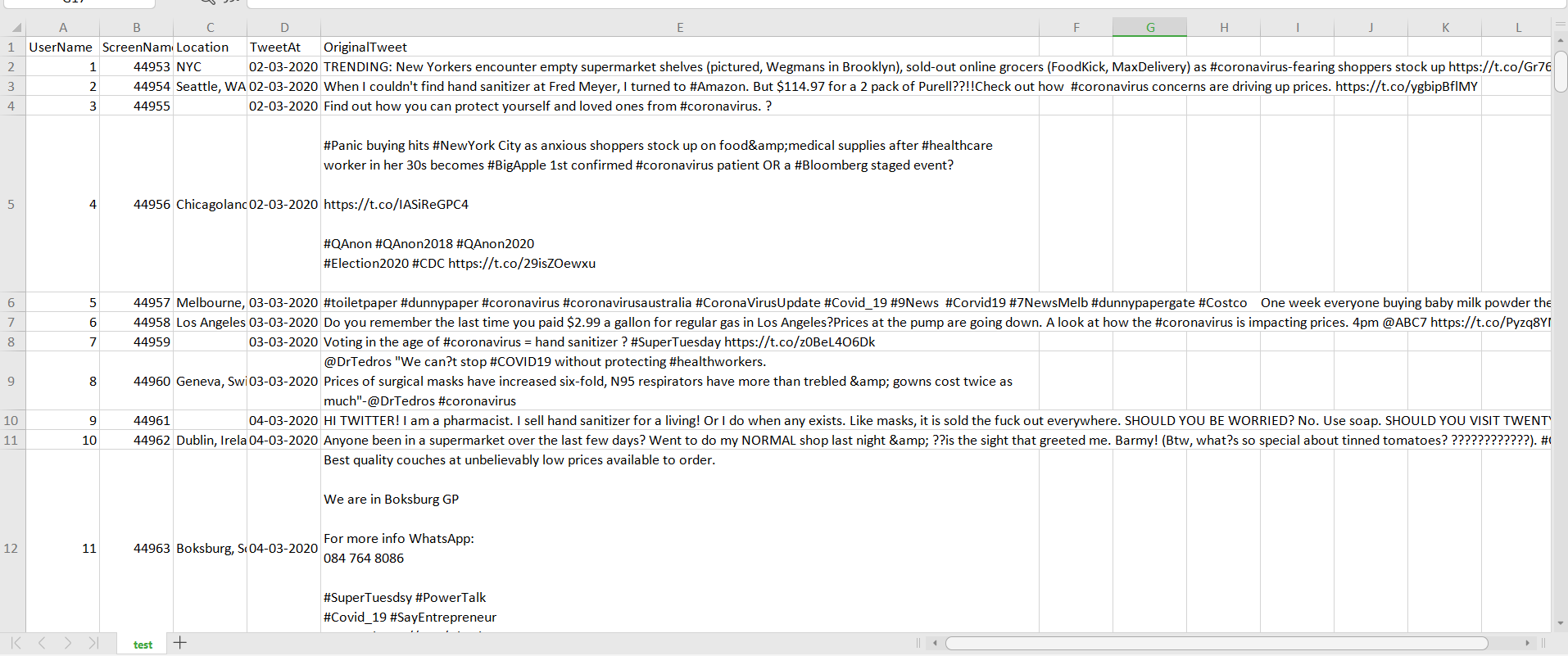
****

****

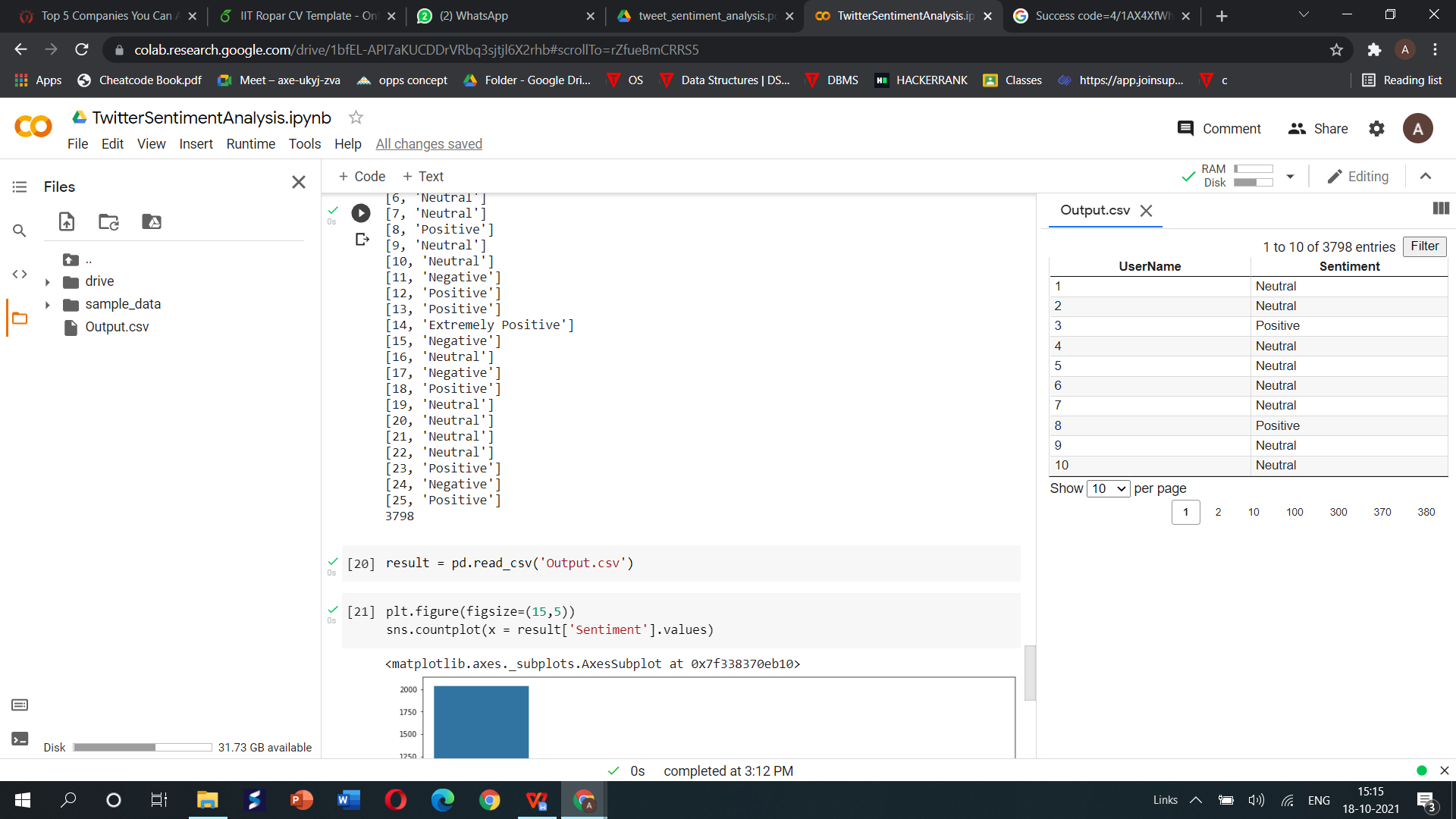
**TRAINING DATA:**

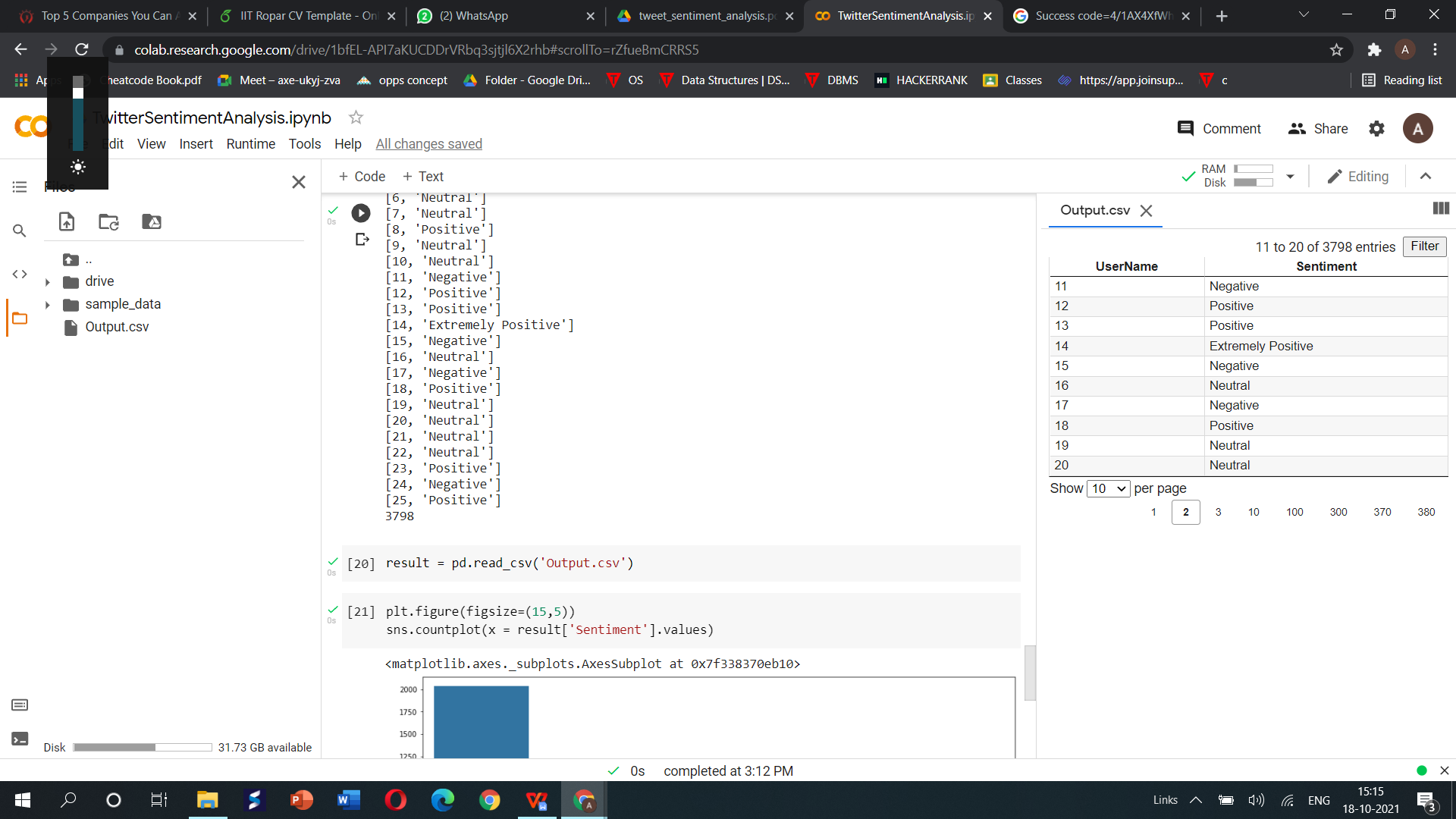
****

**TEST DATA:**

****

**OUTPUT CSV FILE:**

****

****

**LOGIC USED IN ABOVE IMPLEMENTATION:**

Initially, train and test datasets are uploaded in a Google drive. Then, Google Colab is connected (or mounted) with Google drive to import datasets. Google Colab is a specialized version of the Jupyter Notebook. It runs on the cloud and offers free computing resources. After importing dataset from drive, stop words from all the texts present in the dataset are removed. Stop words are words which do not give much meaning to a sentence and they can be removed before modelling tasks. NLTK package supports stop word removal.

After that a very important task called **data cleaning** is done. Data cleaning is a process of removing symbols and non-alphanumeric character and entire text is converted to lower case. Tokenization is carried out which is a way of separating a piece of text into small units called tokens. Finally, all the tweets have been converted to corpus. Then, the cleaned data needs to be converted into a numerical format where each word is represented by a matrix. This is done by TF-IDF method.

Logistic regression model predicts an outcome for a binary value from one or more response variables. Where Random forest algorithm predicts based on combination of different decision trees. Accuracy, recall and precision values, confusion matrix of both the models are found after implementing both the algorithms. By considering these resultant values, we can conclude that logistic regression shows higher performance and more accuracy in prediction than random forest.

Final classification is done using logistic regression model and the final result is stored in output.csv file.