**Data Science Project**

**Topic : Twitter Sentiment Analysis of Product Reviews**

**Team Members:**

Anushka Tadphale - 111903019

Apurva Deshpande - 111903020

Ashlesha Joshi - 111903022

1. **Introduction:**

Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source. Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of the people about a variety of topics.

1. **Methodology Used:**

* Data Collection: Dummy dataset was created using web scraping of tweets using selenium and TweePY API. Due to limitations on number tweets that can be extracted through scraping, we had to take into consideration the pre-collected dataset.

The dataset used in this project is taken from Data World Crowd-flower

<https://data.world/crowdflower/apple-twitter-sentiment>

Attributes present in the dataset are:

\_unit\_id: ID for twitter query

\_golden, \_unit\_state, \_trusted\_judgments, \_last\_judgment\_at: related to query.

sentiment: On a scale 0f 1-3-5

sentiment:confidence: confidence level for given sentiment value

date: date of tweet

id: if of twitter query

query : #AAPL or @apple

sentiment\_gold: possible sentiment values

text: text content of tweet

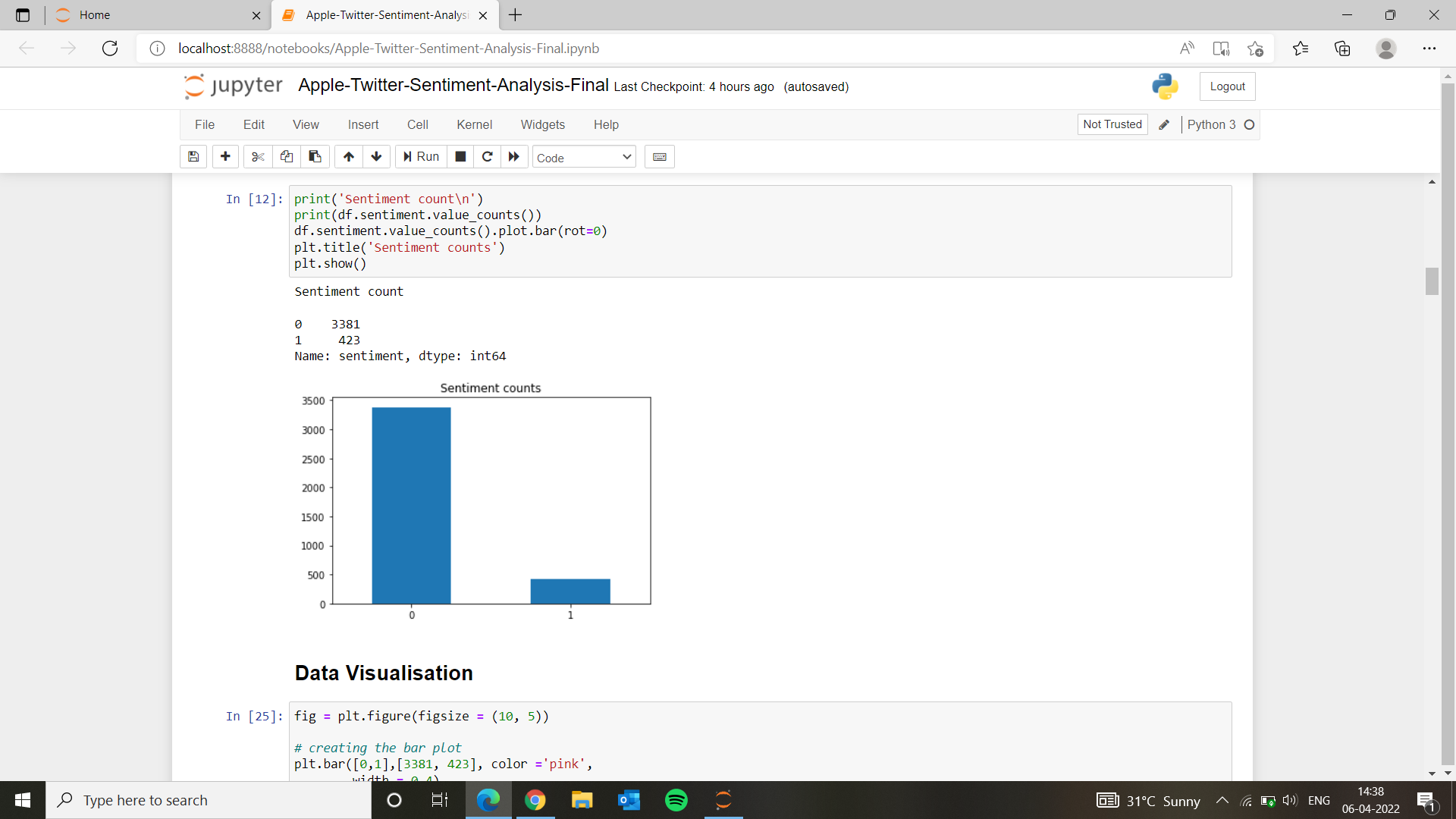
* Data Preprocessing:

Following preprocessing steps were followed on the data set.

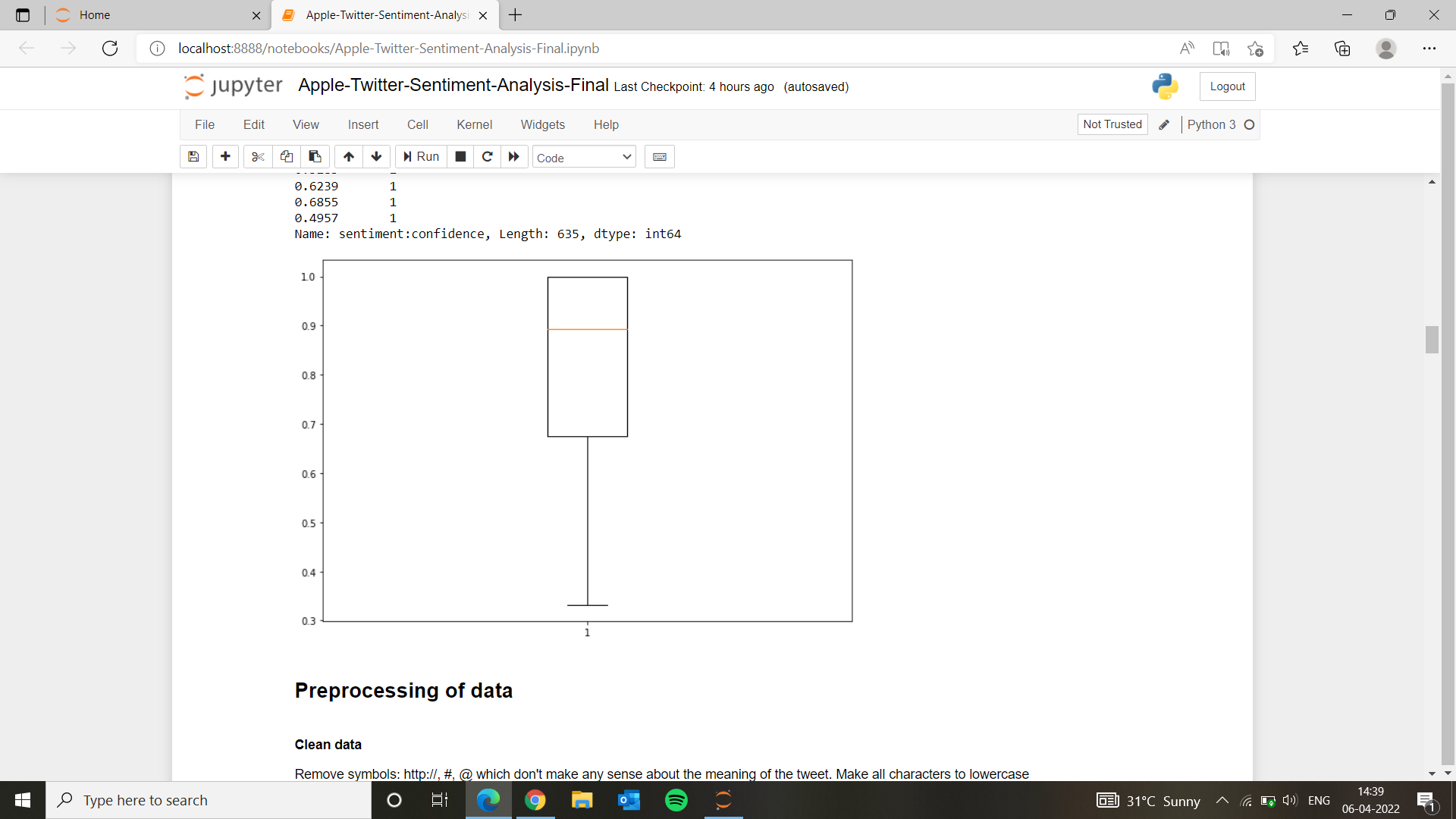
1. Removing unnecessary data fields: few columns like id(same for all tweets), date of tweet, golden, etc are irrelevant for sentiment analysis. So they are dropped.
2. Values of sentiments 1-3-5 are encoded to 0(negative) and 1(positive)
3. Clean data: Remove symbols: http://, #, @ which don't make any sense about the meaning of the tweet. Make all characters lowercase. Remove stop words and words shorter than length 3
4. Tokenization: Split into words
5. Stemming: reducing word to their root form
6. Create a bag of words and split dataset into training and testing
7. Vectorize with TFIDF

* Data Visualization:

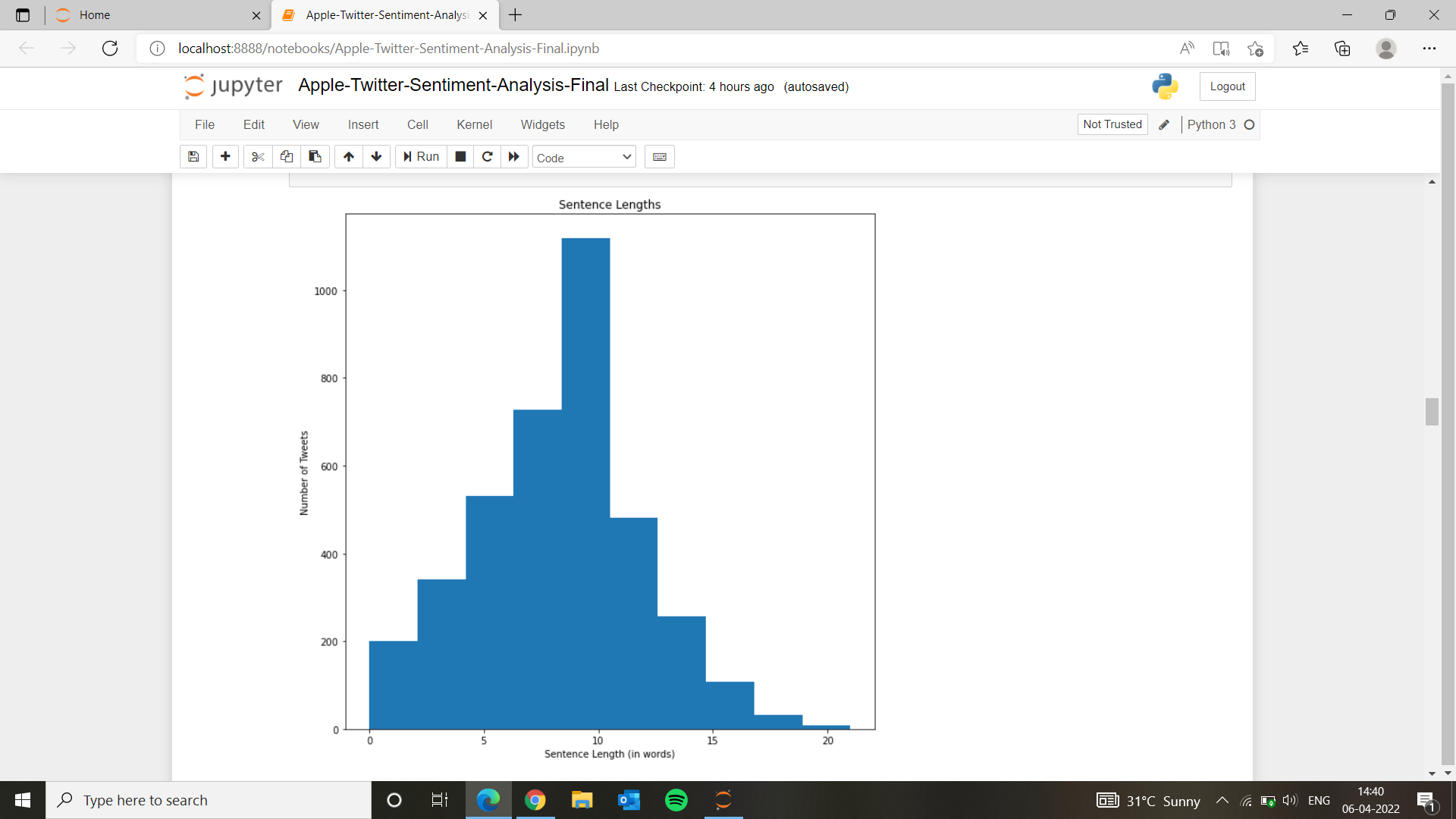
1. Distribution of tweets with positive and negative sentiments



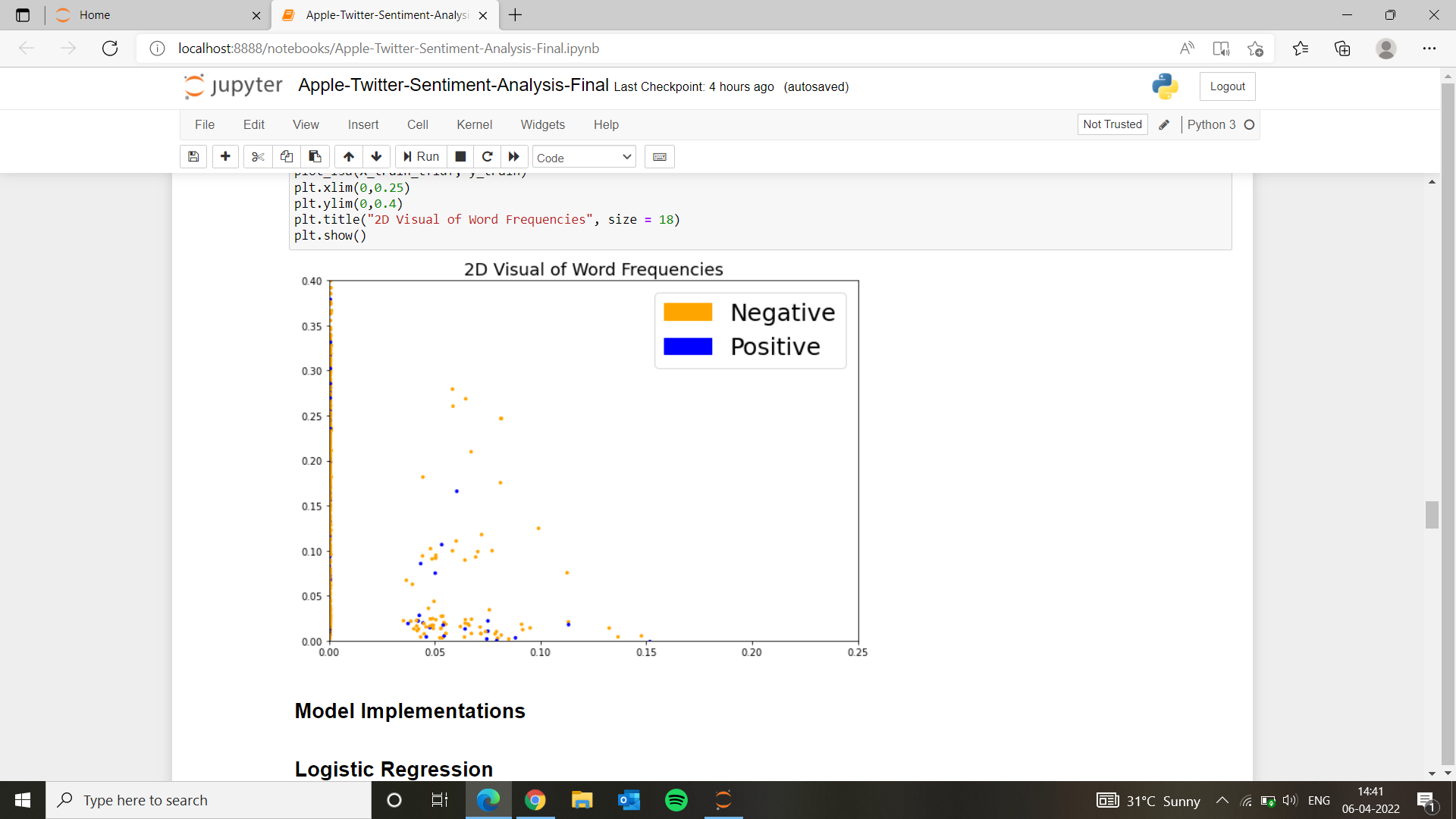
1. Box plot of sentiment value confidence



1. Sentence lengths:



1. Visualization of frequency of words in sentences



* Model Implementation:

Machine learning Model is implemented using 4 Algorithms :

Logistic Regression :

Logistic regression is used for classification. Logistic regression uses a sigmoid function, which gives a probability 0 to 1. Generally applies to binary classification.It is called supervised because the model predictions are iteratively evaluated and corrected against the output values, until an acceptable performance is achieved.

Random Forest :

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Multinomial Naive Bayes :

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance. A feature's existence or absence has no bearing on the inclusion or exclusion of another feature. It uses Bayes Theorem which estimates the likelihood of occurrence based on prior knowledge of the event's conditions.

XgBoost :

XgBoost stands for Extreme Gradient Boosting, written in C++ which optimizes the training for Gradient Boosting. In gradient boosting, each predictor corrects its predecessor’s error.In this algorithm, decision trees are created in sequential form. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. It can work on regression, classification, ranking, and user-defined prediction problems.

* Testing the Model:
* Dataset collected is divided in the ratio 80:20 for Training the model and testing it.
* Models were tested on testing dataset and classification reports were obtained, which shows -

Accuracy of the model

Precision in predicting positive and negative emotions

macro average , weighted average of predicted values are obtained

* **Tools & Libraries Used:**
* Tweepy API - for Web scraping
* Pandas, Numpy, nltk libraries - for Data Preprocessing
* Matplotlib, Seaborn libraries - for Data Visualization
* Scikit learn - for implementing Machine learning algorithms, and testing models
* **Conclusion:**
* Accuracy of models implemented:

| Algorithm | Accuracy |
| --- | --- |
| Logistic Regression | 88.56% |
| Random Forest | 89.09% |
| Multinomial Naive Bayes | 84.88% |
| XG Boost | 90.60% |

* **References**:
* <https://www.researchgate.net/publication/334845233_Real_Time_Twitter_Sentiment_Analysis_for_Product_Reviews_Using_Naive_Bayes_Classifier>
* <https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/>
* <https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/>
* <http://www.iosrjen.org/Papers/Conf.SICTIM-2019/Volume-1/5.%2022-25.pdf>
* <https://www.revuze.it/blog/sentiment-analysis-using-product-review-data/>