

Project Report on

**Twitter Sentiment Analysis**

Submitted in partial fulfillment for the award of **Post Graduate Diploma in Big Data Analytics** from **C-DAC KHARGHAR**

**Guided by:**

**Mr. Pankaj Khambire**

Presented by:

1. **Mr. Jay Sagar PRN : 220340325017**
2. **Ms. Renu Barale PRN : 220340325039**
3. **Mr. Hitesh Kamlikar PRN : 220340325016**
4. **Mr. Harshal Kamble PRN : 220340325015**
5. **Mr. Avdhut Gadhe PRN : 220340325010**
6. **Mr. Mohit Patil PRN : 220340325026**

**Centre of Development of Advanced Computing (C-DAC), Pune**



CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

**This is to certify that**

**Mr. Jay Sagar**

**Ms. Renu Barale**

**Mr. Hitesh Kamlikar**

**Mr. Harshal Kamble**

**Mr. Avdhut Gadhe Mr. Mohit Patil**

# have successfully completed their project on

**Twitter Sentiment Analysis**

## under the guidance of Mr. Pankaj Kambire

**Project Guide Project Supervisor**

**Mr. Pankaj Kambire Mr. Pradeep Tripathi**

## Course Coordinator

**Mrs. Vineeta Singh**



ACKNOWLEDGEMENT

This project “**Twitter Sentiment Analysis**” was a great learning experience for us and we are submitting this work to CDAC Kharghar.

We all are very glad to mention the name of *Mr. Pankaj Kambire* for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to Mr. Pradeep Tripathi for his guidance and support whenever necessary while doing this course Post-Graduate Diploma in *Big Data Analytics (PGDBDA)* through C-DAC Kharghar.

Our most heartfelt thanks go to *Mrs. Vineeta Singh* (Course Coordinator, *PGDBDA*) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility, and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Kharghar.

From:

1. Mr. Jay Sagar (220340325017)
2. Ms. Renu Barale (220340325039)
3. Mr. Hitesh Kamlikar (220340325016)
4. Mr. Harshal Kamble (220340325015)
5. Mr. Avdhut Gadhe (220340325010)
6. Mr. Mohit Patil (220340325026)



**TABLE OF CONTENTS:**

1. Abstract
2. Introduction and Overview of Project
3. Dataset Description
4. Information on Techniques used
5. AWS EMR
6. Data Preprocessing
7. Model Building
8. Prediction
9. Data visualization
10. Conclusion
11. Future scope
12. References



# Abstract

Twitter is a micro-blogging website that allows people to share their views about topics or post messages. There has been a lot of work on the sentiment analysis of Twitter data. This Project involves the classification of tweets into two main sentiments: Positive and Negative. Social media websites have emerged as one the platforms to raise users’ opinions and influence the way any business is commercialized.

People's opinions matter a lot to analyze how information propagation impacts life in a large-scale network like Twitter. Sentiment Analysis of the tweets determines the polarity and inclination of a vast population toward a specific topic, item, or entity. The application of such analysis can be easily observed during public elections, movie platforms, brand endorsements, and many other fields. Spurred by that growth, companies, and media organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services.

In this project, we exploited the fast and in-memory computation framework 'Apache Spark' to extract tweets and perform sentiment analysis. The primary aim is to provide a method for analyzing sentiment scores in noisy Twitter streams. This paper reports on the design of sentiment analysis, extracting a vast number of tweets. Results classify users’ perceptions via tweets into positive and negative. Secondly, we discuss various techniques to carry out a sentiment analysis on Twitter data in detail.



# Introduction and Overview of Project

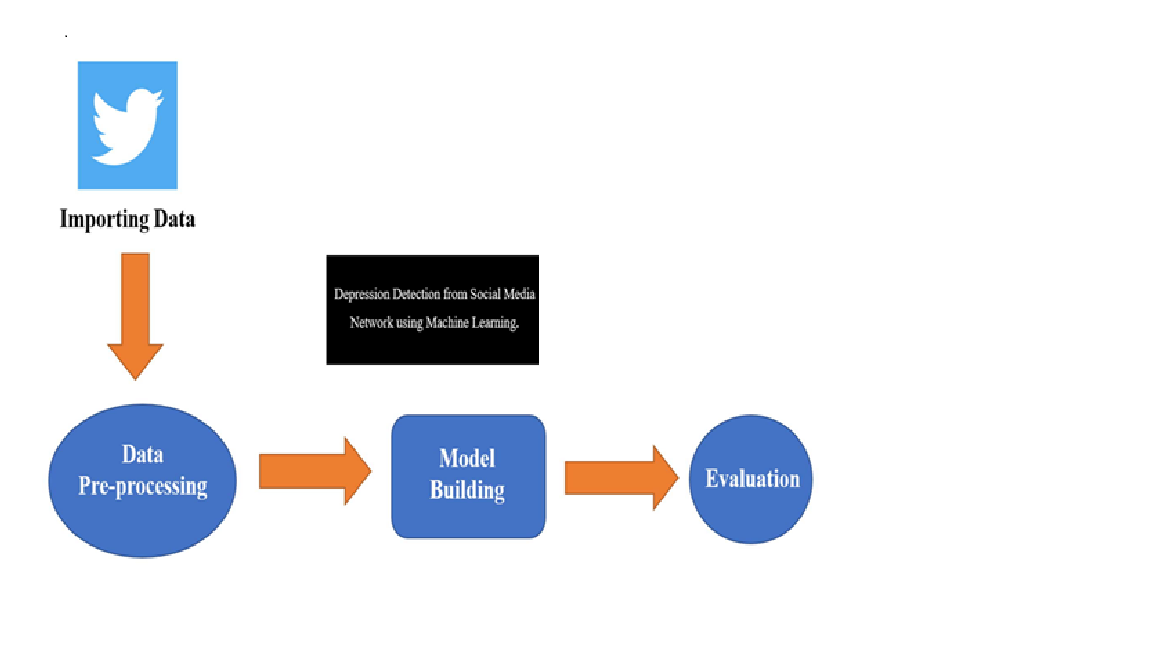
As the internet is growing bigger, its horizons are becoming wider. Social Media and Microblogging platforms like Facebook and Twitter dominate in spreading encapsulated news and trending topes across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These topics are generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements, and entertainment like movies and award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further helps in enhancing marketing strategies. One such example can be leaking pictures of upcoming iPhones to create hype to extract people's emotions and market the product before its release. Thus, there is a huge potential for discovering and analyzing interesting patterns from the infinite social media data for business-driven applications. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions, and emotions expressed within an online mention. The retention is to gain an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations Ito positive, negative, or neutral labels.

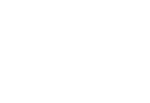
Many people use social media sites for networking with other people and to stay up-to-date with news and current events. These sites (Twitter, Facebook, Instagram, and google+) offer a platform for people to voice their opinions. For example, people quickly post their reviews online as soon as they watch a movie and then start a series of consents to discuss the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, and rate the performance of not only any movie but about other products and to know whether it will be a success or not.



This type of vast information on these sites can be used for marketing and social studies. Therefore, sentiment analysis has wide applications and includes emotion mining, polarity, classification, and influence analysis.

Moreover, extensive usage of slang words, acronyms, and vocabulary words are quite common while tweeting online. The categorization of such words per polarity gets tough for natural processors involved. In this project, we used Apache Spark's fast processing capabilities to analyze sentiment from tweets. Next, the problems we came across and the challenges we resolved during implementation are specified. The further future scope is discussed. Finally, concludes the report.





# Dataset Description

The Data required for this Project has been taken from Kaggle. This is the sentiment 140 dataset. It contains 1,600,000 tweets extracted using the Twitter API. The tweets have been annotated (0 = negative, 1 = positive) and they can be used to detect sentiment.

## Content :

It contains the following 6 fields:

1. Target: the polarity of the tweet (*0* = negative, *1*= positive)
2. Id: The id of the tweet
3. Date and Time: the date of the tweet
4. Flag: If there is no query, then this value is NO\_QUERY.
5. User: the user that tweeted.
6. Text: the text of the tweet



# 4. Information on Technologies used

## Apache Spark :

## We performed cluster computing and processing in Apache spark. We used Spark SQL module for data abstraction on structured data.

## Python :

We used python based API Pyspark for Apache Spark for this project. Also, we used various libraries like nltk, pandas, and sklearn libraries. We connect AWS service S3 with pyspark for pre-processing and modeling.

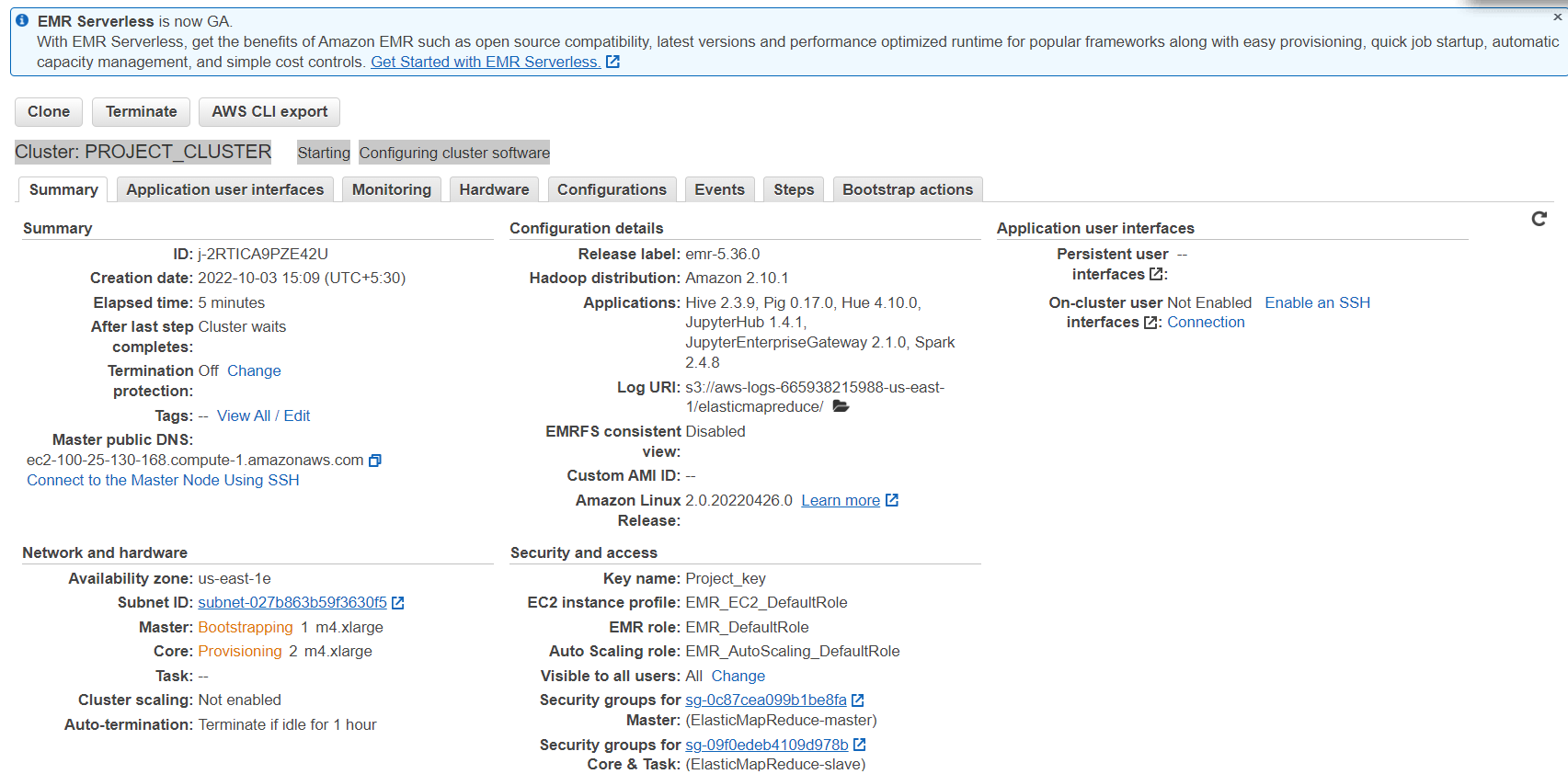
## AWS :

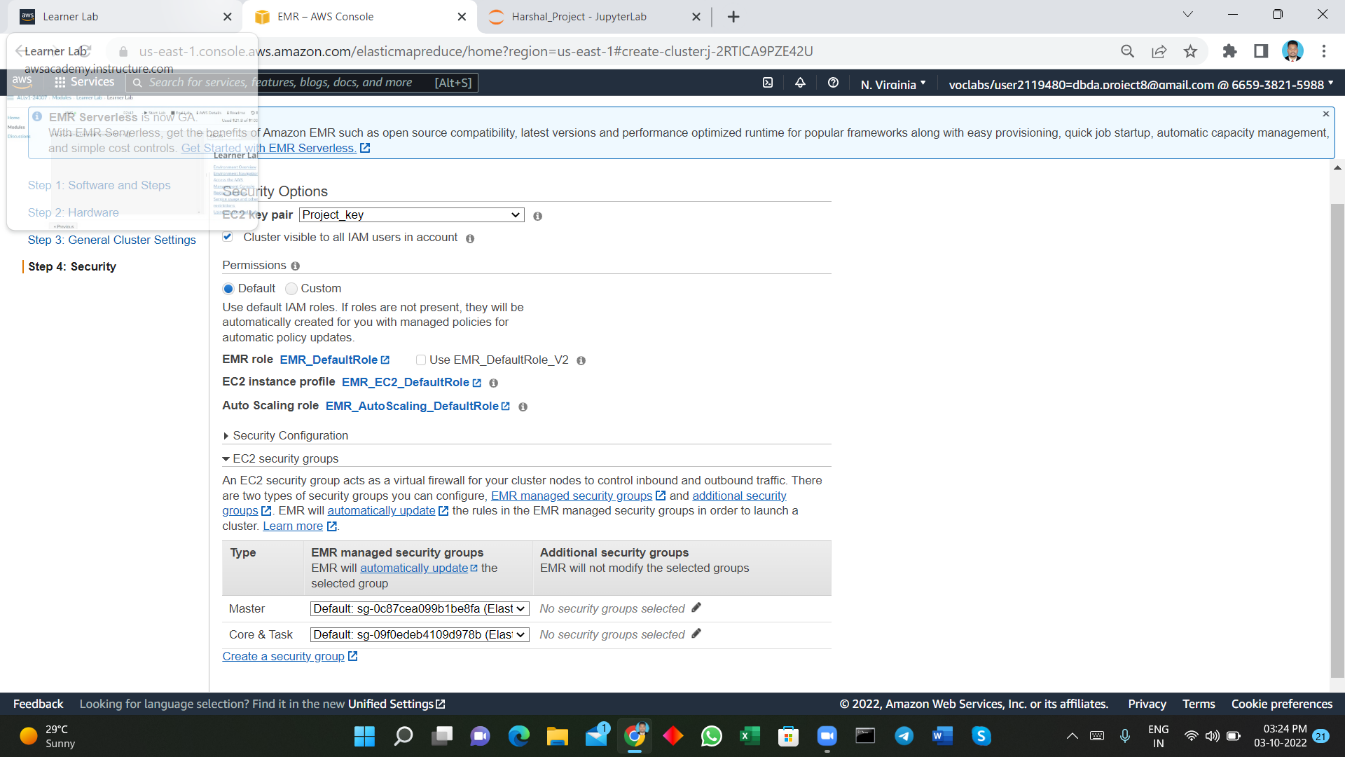
## For this project, we used AWS services. Many services like EMR and EC2 are used for creating clusters and instances. We stored the dataset on S3 for processing. Also, we connect the Jupyter notebook to the EC2 instance for prediction purposes.

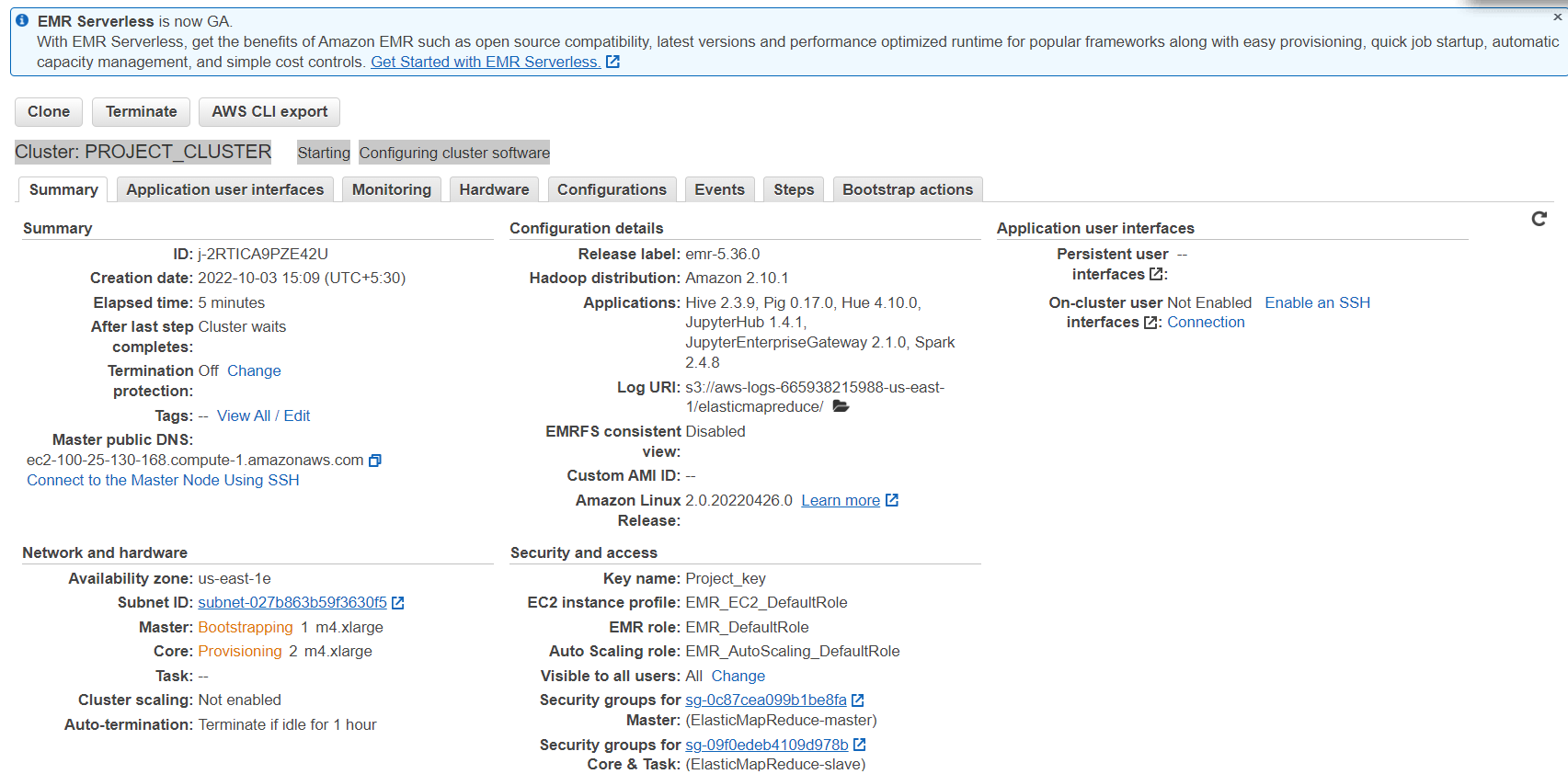
**5. AWS EMR**

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark. EMR is a managed cluster platform that assists organizations in running Big Data frameworks on AWS to analyze and process large sets of data more efficiently.

An EMR notebook is a "serverless" notebook that you can use to run queries and code. Unlike a traditional notebook, the contents of an EMR notebook itself—the equations, queries, models, code, and narrative text within notebook cells—run in a client. The commands are executed using a kernel on the EMR cluster. Notebook contents are also saved to Amazon S3 separately from cluster data for durability and flexible re-use.

****

****

****

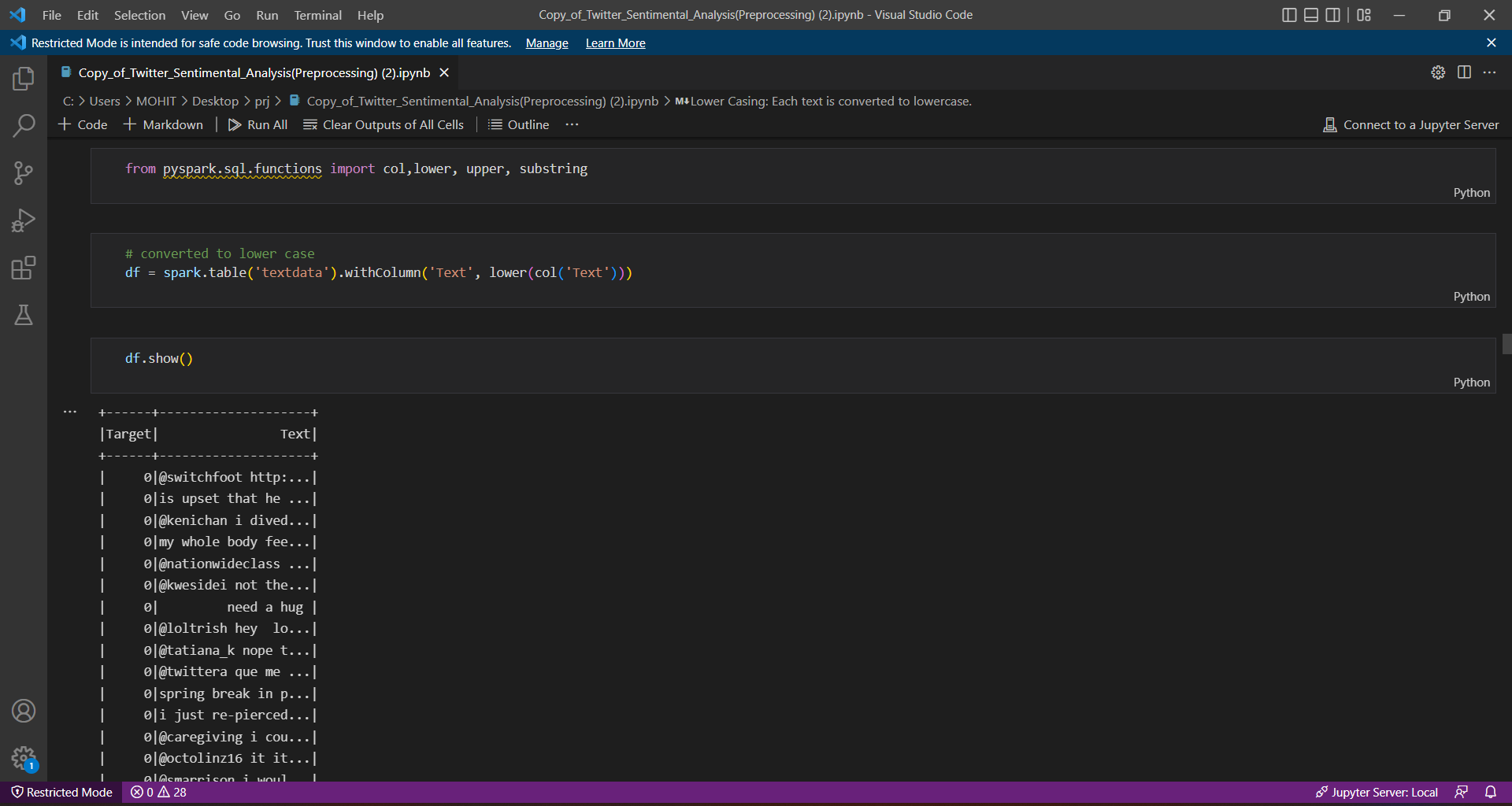


# 

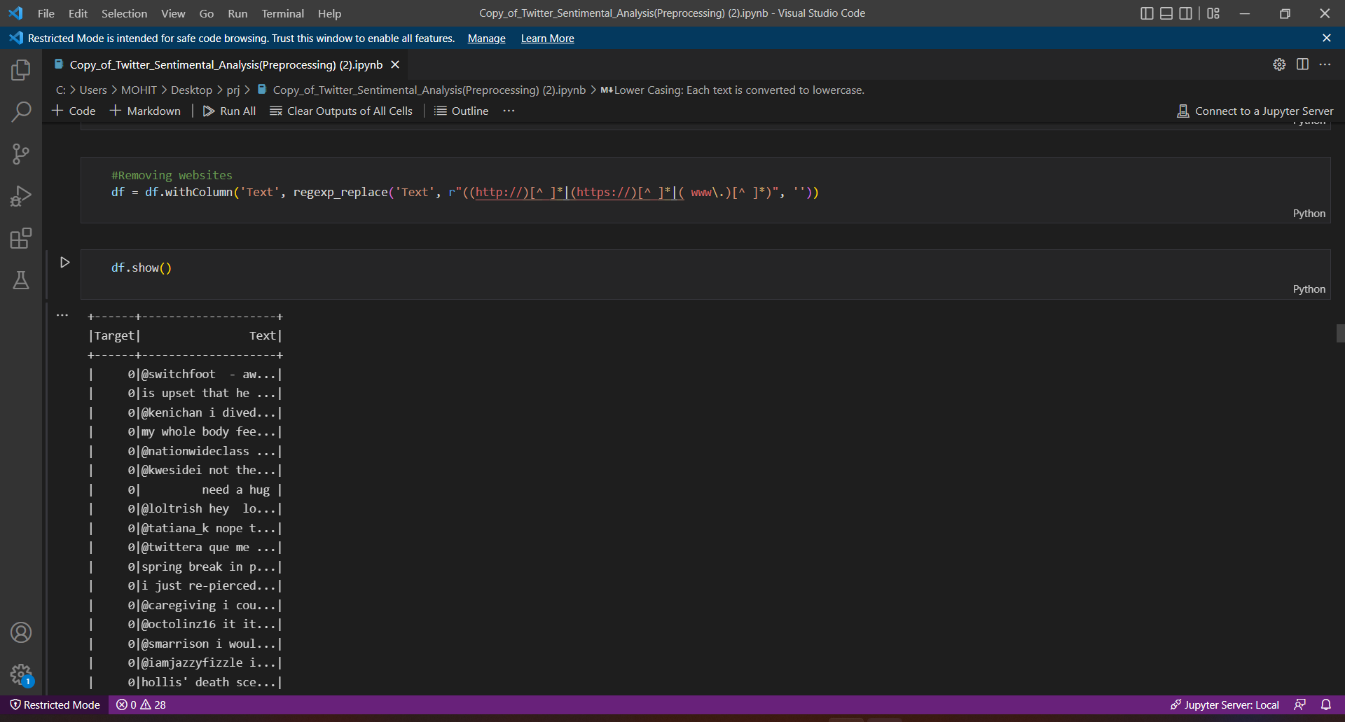
# 6. Data Pre-processing

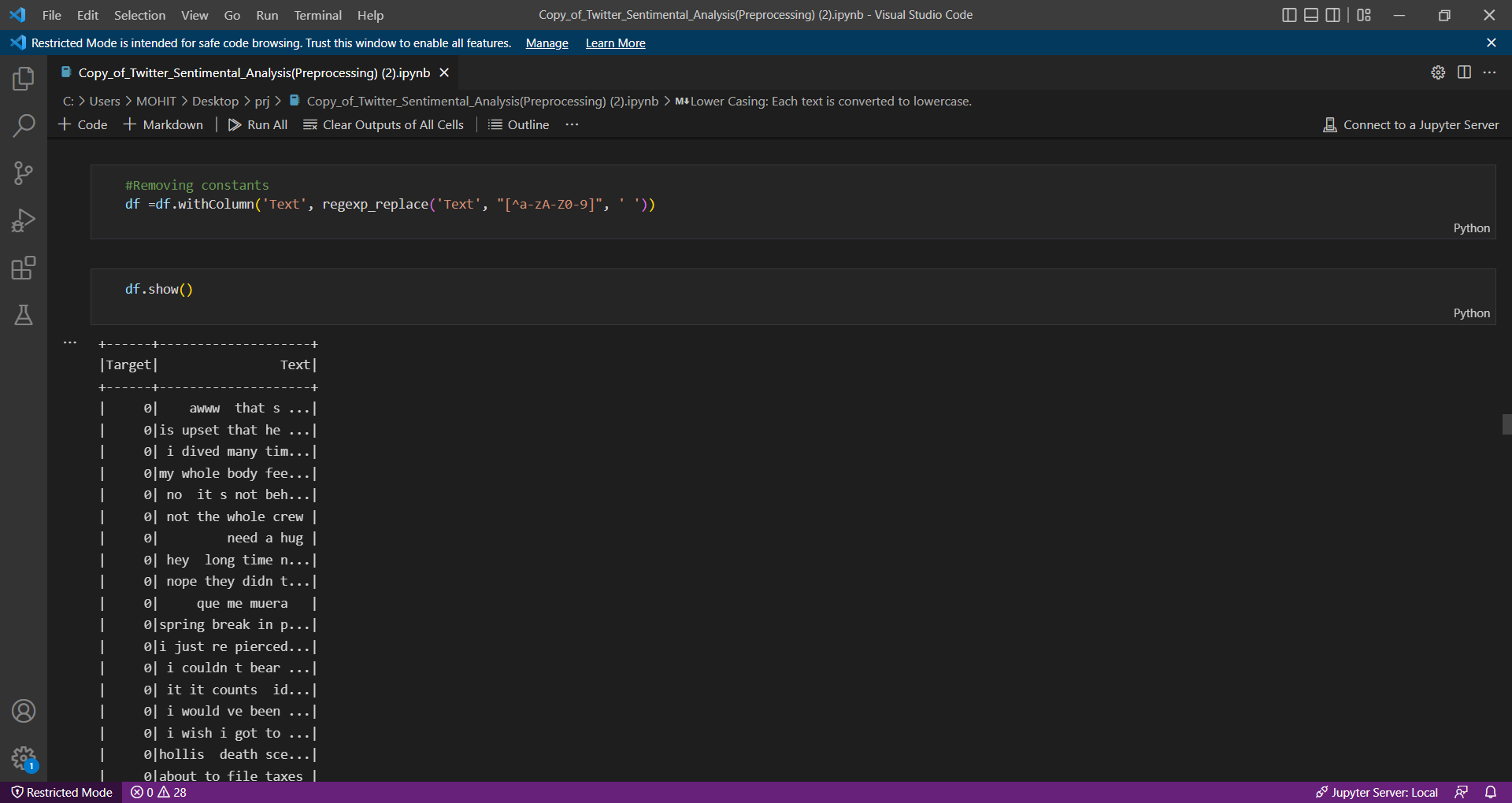
For this project, we have used pyspark for data preprocessing. We perform the following steps on the dataset,

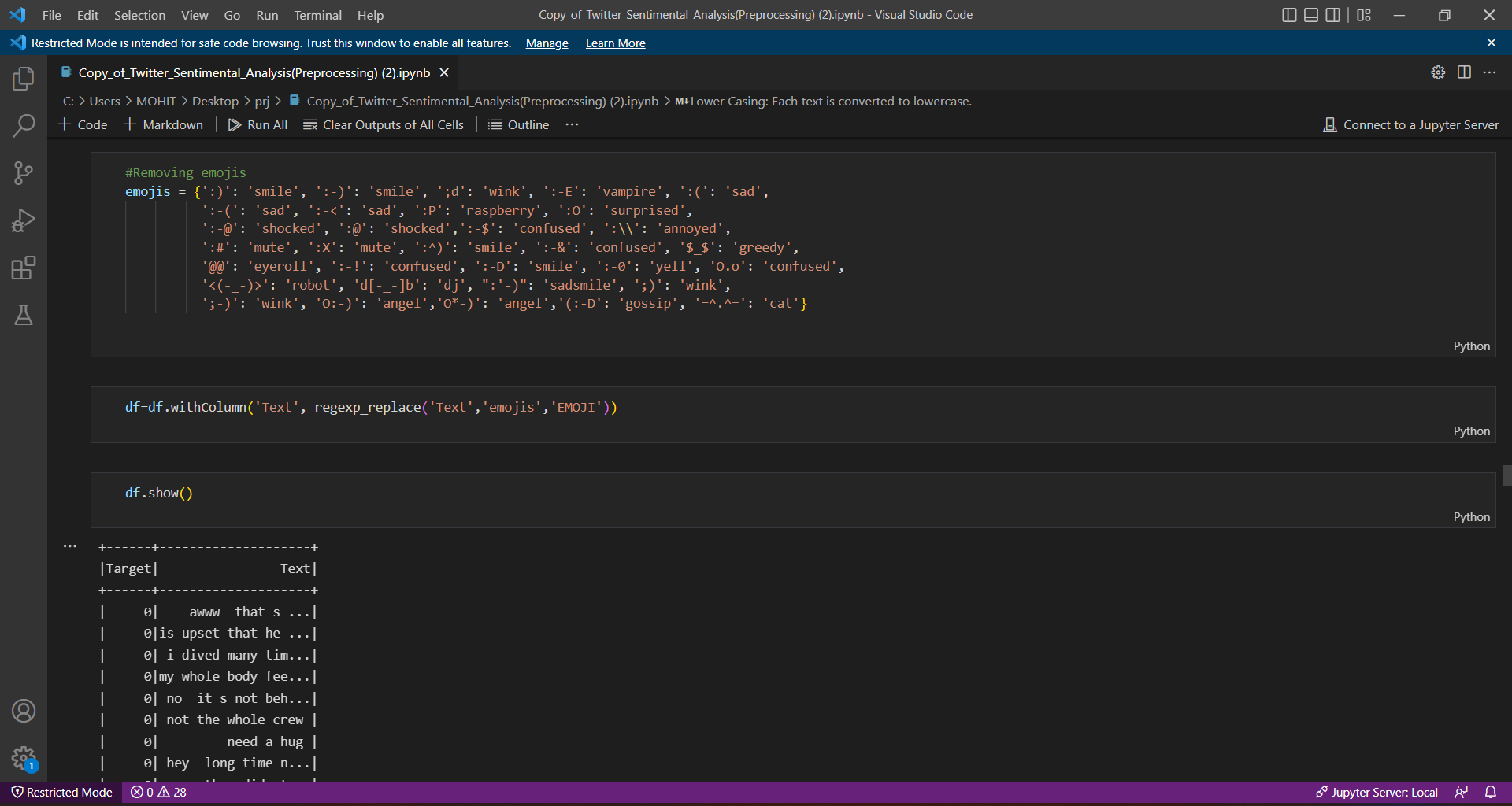
1. Lower Casing: Each text is converted to lowercase.
2. Replacing URLs: Links starting with "http" or "https" or "www" are replaced by "URL".
3. Replacing Emojis: Replace emojis by using a pre-defined dictionary containing emojis along with their meaning. (eg: ":)" to "EMOJIsmile")
4. Replacing Usernames: Replace @Usernames with the word "USER". (eg: "@Kaggle" to "USER")
5. Removing Non-Alphabets: Replacing characters except Digits and Alphabets with a space.
6. Removing Short Words: Words with lengths less than 2 are removed.
7. Removing Stopwords: Stopwords are English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. (eg: "the", "he", "have")
8. Lemmatizing: Lemmatization is the process of converting a word to its base form. (e.g: “Great” to “Good”)









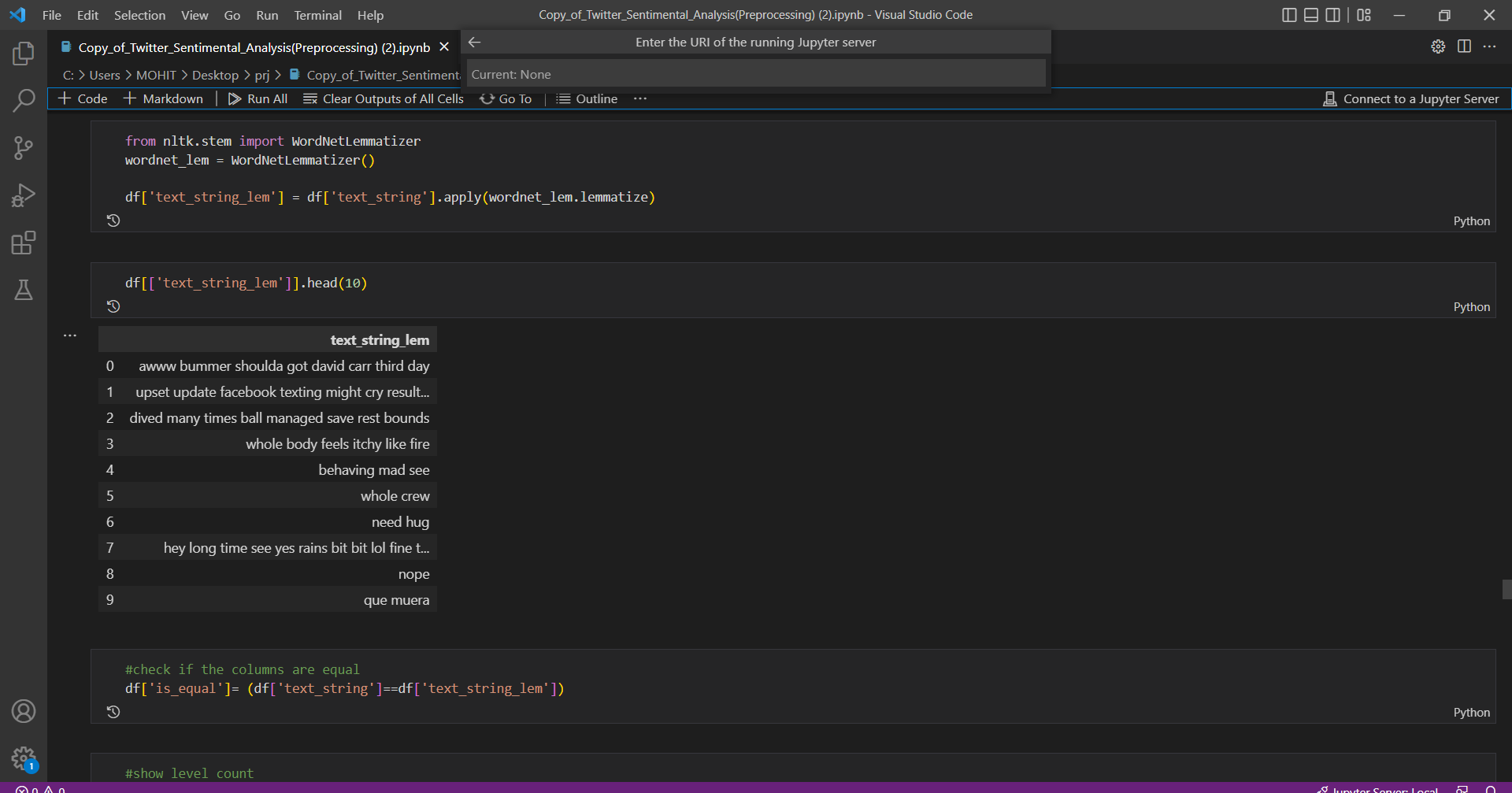




## Lemmatizing:

Wordnet is a large, freely, and publicly available lexical database for the English language aiming to establish structured semantic relationships between words. It offers lemmatization capabilities as well and is one of the earliest and most commonly used lemmatizers.

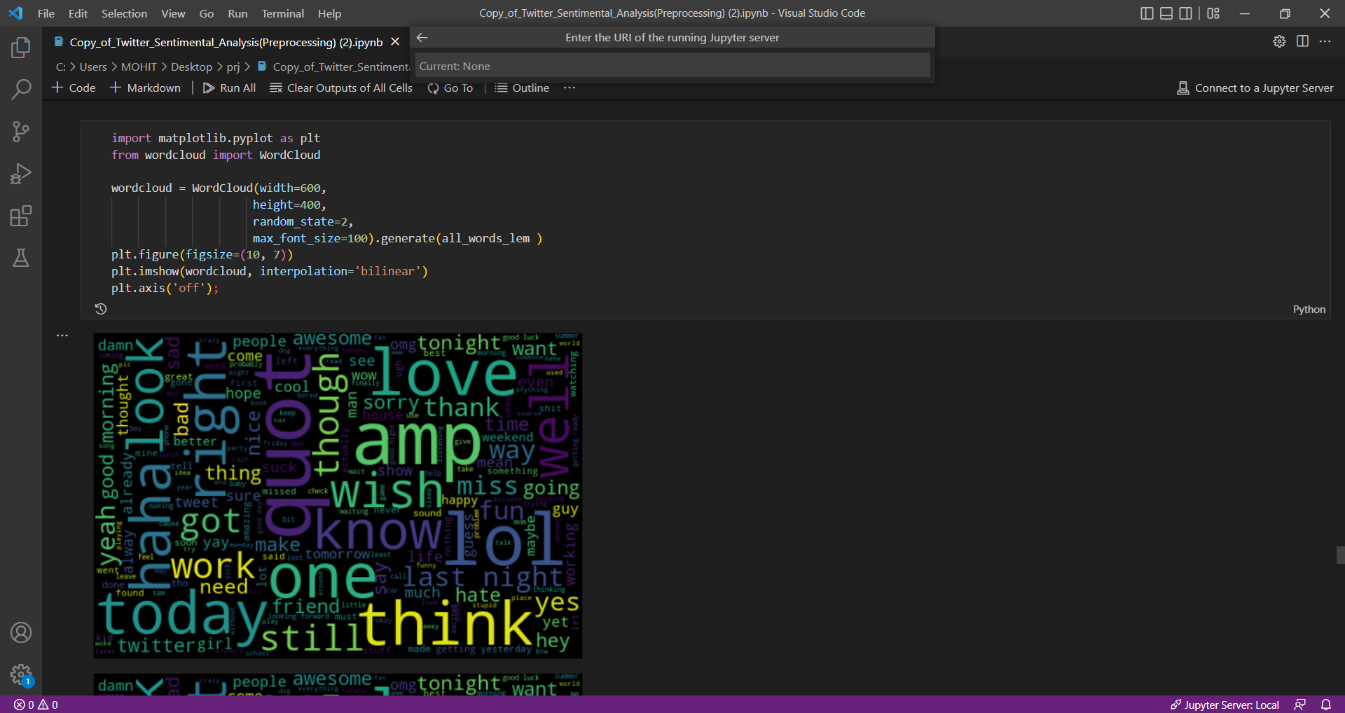
In order to lemmatize, you need to create an instance of the WordNetLemmatizer() and call the lemmatize() function on a single word.





**Analysing the data**:

We analyse the preprocessed data to get an understanding of it. We plot **Word Clouds** for **Positive and Negative** tweets from our dataset and see which words occur the most.



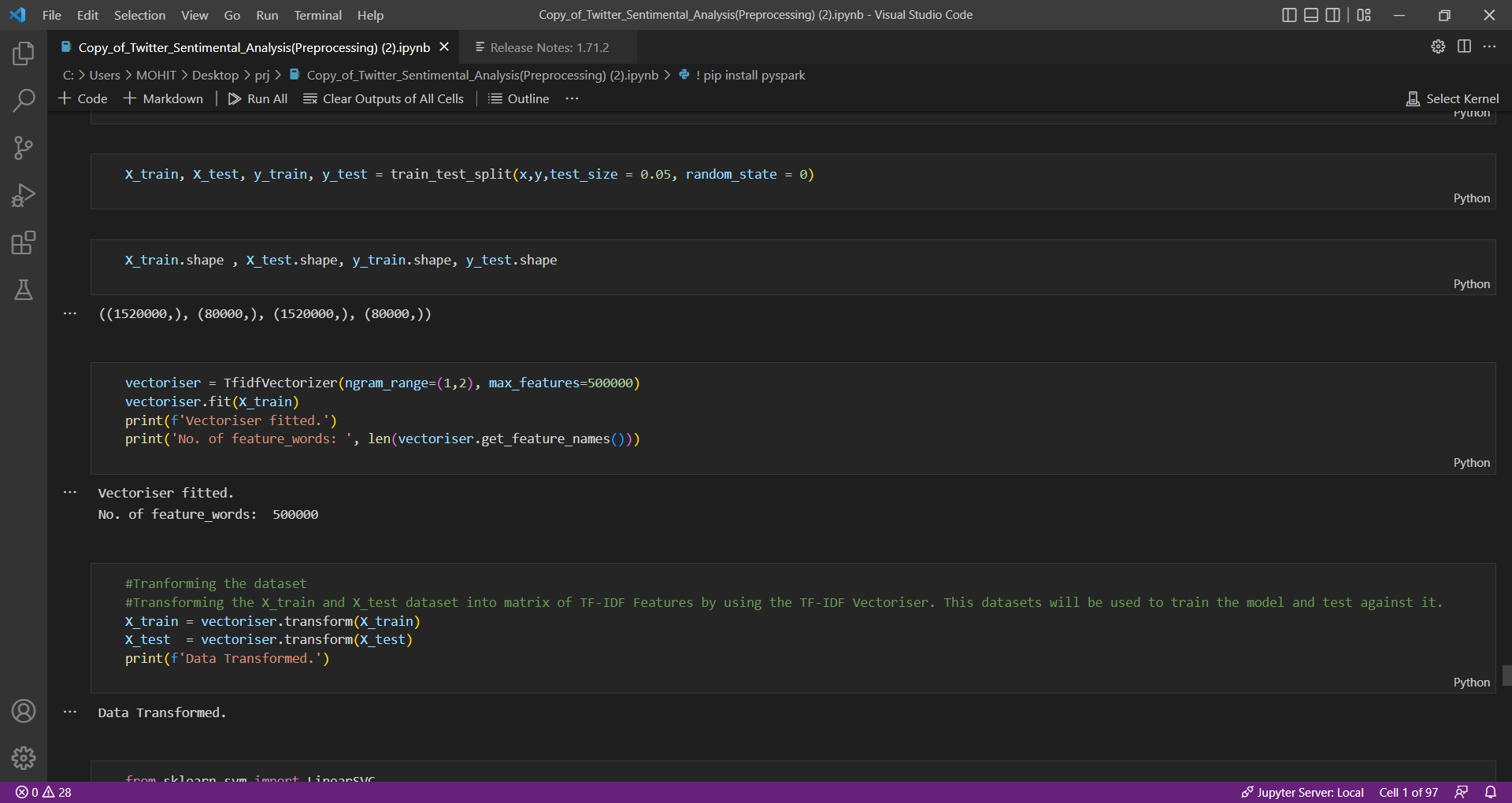
## 

## Splitting the Data:

The Preprocessed Data is divided into 2 sets of data.

Training data: The dataset upon which the model would be trained on. Contains 95% data.

**Test Data:** The dataset upon which the model would be tested against. Contains 5% data.





# 7. Model Building

**Logistic Regression :**

This type of statistical model (also known as the logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voting or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

# Naïve Byes :

The naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. It implements the Bayes theorem for the computation and used class levels represented as feature values or vectors of predictors for classification. Naive Bayes Algorithm is a fast algorithm for classification problems.

# SVM :

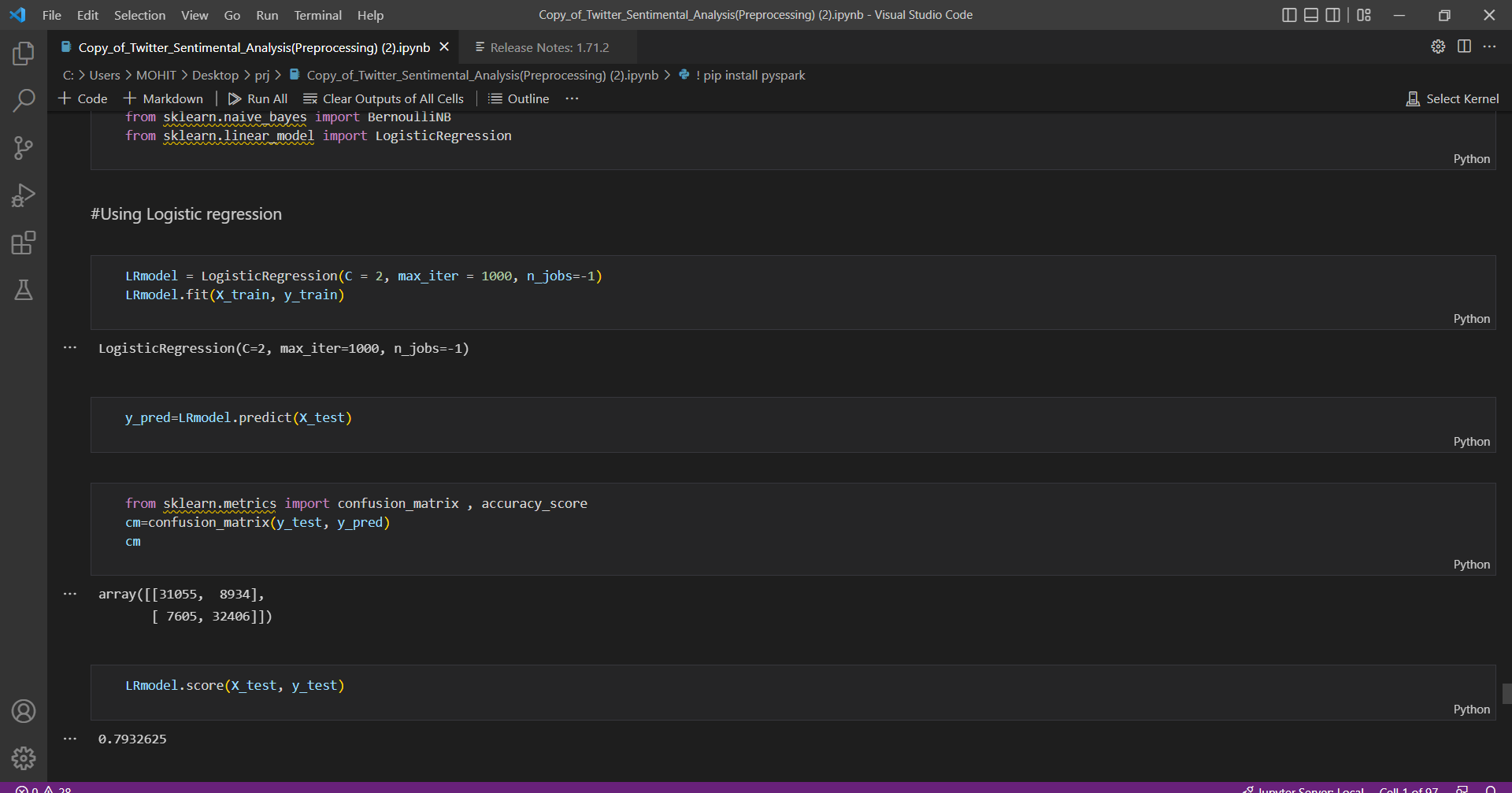
A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they’re able to categorize new text.

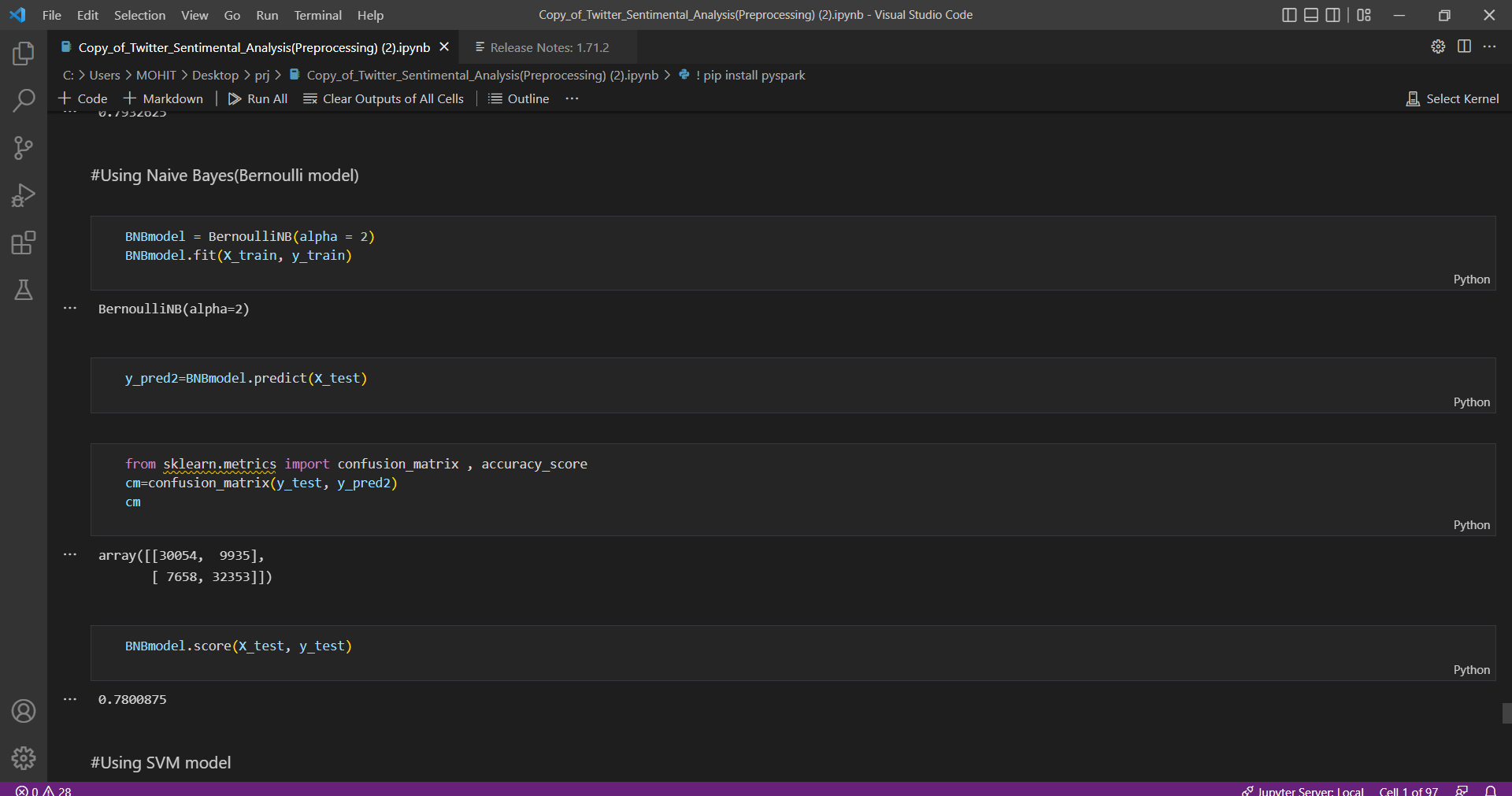
Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands).



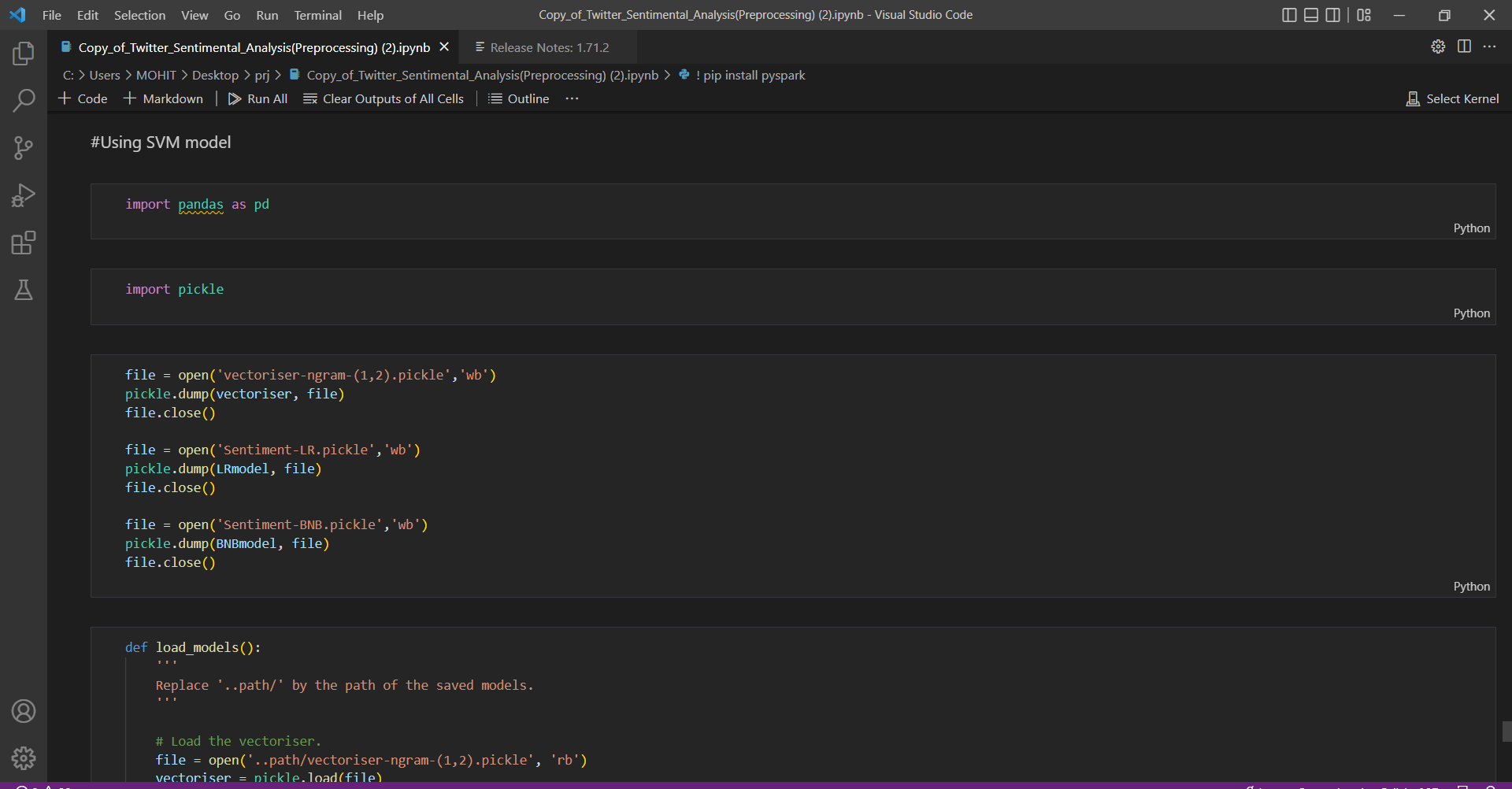
# 8. Prediction

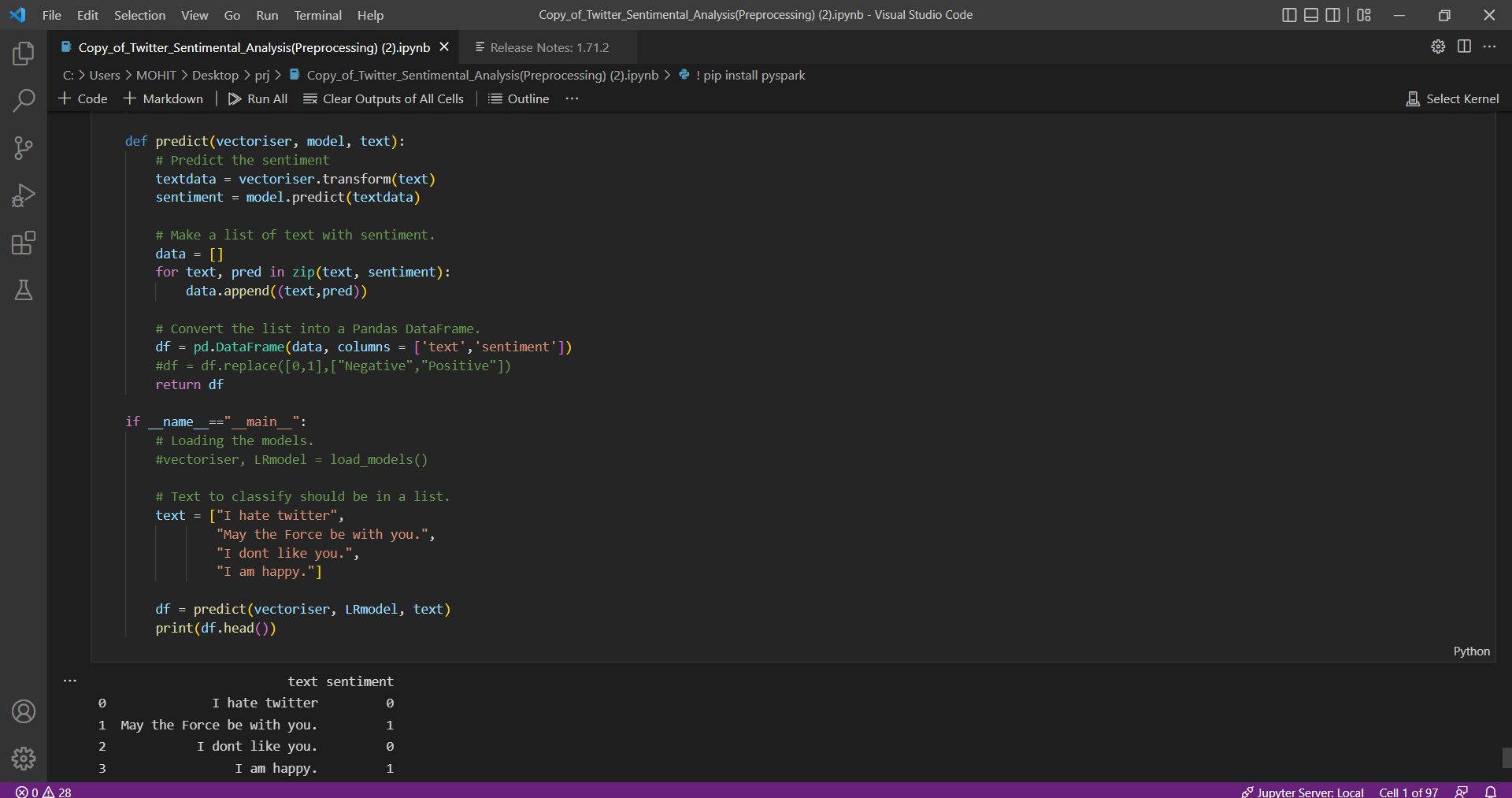
We got pretty decent accuracy. We have to write the tweet here in the model to get the sentiment, probability, and expression related to the particular tweet.













# Data Visualization

# 

We have used Power BI for Data visualization. In that we got to know positive and negative sentiments count of users. We have also got the count of words occurring more frequently in raw data as well as processed data. Below are the various visual insights we have plotted through data.

# 

# 

# 10.Conclusion

Twitter is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. Apache Spark proved prolific in extracting live streams of data and has the further capability to store batches of data in HDFS and other major conventional storages. The processing capabilities of Spark make the project flexible to further extend to multiple nodes, thereby supporting distributed computing. Real-time data analysis makes it possible for business organizations to keep track of their services and generates opportunities to promote, advertise and improve from time to time.



# 11. Future Scope

By examining and evaluating customer sentiments with such tools, brands can gain a comprehensive understanding of consumer behaviors and, as a result, better serve their audiences with the products, services, and experiences they offer. However, sentiment analysis will delve deeper in the future, beyond the concept of positive, negative, or neutral, to reach and comprehend the significance of understanding conversations and what they reveal about consumers.

As a result, sentiment analysis is becoming more important for these businesses as the data underlying those interactions grows larger and more complex.

Even today, corporations and brands perform the vast majority of sentiment analysis in any project, utilizing data from social media, survey responses, and other sources of user-generated content.

Furthermore, brands will be able to easily customize and personalize their services with a thorough understanding and a much larger and more comprehensive database. Instead of segmenting markets based on age, gender, income, and other surface demographics, businesses can segment even further based on how their audience members actually feel about the brand.

A report from Market Research Future claims that the global Sentiment Analytics market is expected to grow to USD 6 billion by 2027. With its ability to deliver invaluable insights about customer behavior and target audience persona, Sentiment Analysis is sure to be a part of the Twitter marketing strategy of corporates for the foreseeable future.



# 12. Bibliography

1. <https://louisazhou.gitbook.io/notes/spark/data-cleaning-with-apache-spark>
2. <https://www.kirenz.com/post/2021-12-11-text-mining-and-sentiment-analysis-> with-nltk-and-pandas-in-python/text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/
3. [Overview of Amazon EMR - Amazon EMR](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html)