****

**Facebook Data Analysis**

Project Report

Submitted as part of the course

Big Data Frameworks - CSE3120

School Of Computer Science and Engineering

VIT Chennai

Winter 2021-2022

Course Faculty:Dr. Amrit Pal

**Submitted By Group - Flink**

*· Nalla Vedavathi - 19MIA1106*

*· Sneha. M - 19MIA1101*

*. Samyuktha Reddy - 19MIA1058*

*. Bachu Akshitha - 19MIA1096*

**ACKNOWLEDGEMENT:**

Primarily, we would like to thank God for giving us the resources and strength for being able to complete this project with success.

Further, we would like to express my special thanks and gratitude to our Big Data Frameworks Faculty Dr. Amrit Pal, whose valuable guidance has helped us patch this project and make it full proof success. His suggestions and instructions have served as the major contributor towards the completion of this project. Throughout the course, he has always entertained and shared his knowledge on this topic with great enthusiasm. We would like to thank him for his constant support and encouragement towards making this project.

We would also take this opportunity to thank Dr. Jagadish Kannan R., Dean School of Computer Science Engineering (SCOPE), for extending the facilities of the school towards our project and for his unstinting support.

Finally, we would like to thank our parents and family who have encouraged us throughout this project to come up with new ideas and helped us get through problems with valuable solutions.

We have learnt a lot through this project and feel ourselves prepared to give solutions for upcoming challenges towards Facebook Data Analysis.

**Table Of Contents:**

|  |  |  |
| --- | --- | --- |
| ***s.no*** | ***Title*** | ***Page Number*** |
| ***1*** | **Title Page** | **3** |
| ***2*** | **Abstract** | **3** |
| ***3*** | **Introduction** | **4** |
| ***4*** | **Proposed System** | **5** |
| ***5*** | **Explanation Of the Project Components** | **6** |
| ***6*** | **Deployment** | **8** |
| ***7*** | **Results** | **46** |
| ***8*** | **Conclusion And Future Work** | **48** |
| ***19*** | **References** | **49** |

**FACEBOOK DATA ANALYSIS:**

In this Project we are going to Analyze the Facebook data using Spark for the purpose of better decision making the business.

**Abstract:**

The concept of big data has been around for years; most organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get significant value from it. But even in the 1950s, decades before anyone uttered the term “big data,” businesses were using basic analytics (essentially numbers in a spreadsheet that were manually examined) to uncover insights and trends.

The new benefits that big data analytics brings to the table, however, are speed and efficiency. Whereas a few years ago a business would have gathered information, run analytics and unearthed information that could be used for future decisions, today that business can identify insights for immediate decisions.

The ability to work faster – and stay agile – gives organizations a competitive edge they didn’t have before. Through our project we intend to carry out analysis on a preferably large dataset. So we have chosen the dataset obtained from several Facebook users. By carrying out certain operations, we intend to harness their data and use it to identify new opportunities.

Through our project we intend to carry out analysis on a preferably large dataset. So we have chosen the dataset obtained from several Facebook users. By carrying out certain operations, we intend to harness their data and use it to identify new opportunities.

The main Objective of the project is to analyze the Facebook data using Hadoop for the purpose of better decision making for the business. Business owners utilize the data to understand customer needs and their behavior to make profit in their business. Facebook data analysis is the process of collecting, analyzing Facebook data and visualizing extracted results to the end users.

**Introduction:**

Smartphones without social media usage in the daily lifestyle of people is unthinkable. As per 2017 statistics, nearly 1.37 billion daily active users for Facebook. Every user contributes some type of data in structured or semi-structured or unstructured data format.

Business owners utilize this data to understand customer needs and their behavior to make profit in their business. Facebook data analysis is the process of collecting, analyzing Facebook data and visualizing extracted results to the end user.

It’s important to be able to analyze customers and their behavior on a micro level due to Facebook’s ever-changing algorithm, and the implications for our content and business.

Facebook is said to have more than 500 million users in 2010. The field of social networks and their analysis has evolved from graph theory, statistics and sociology and it is used in several other fields like information science, business application, communication, economy etc.

Analyzing a social network is similar to the analysis of a graph because social networks form the topology of a graph. Graph analysis tools have been there for decades. But they are not designed for analyzing a social network graph which has complex properties. An online social network graph may be very large. It may contain millions of nodes and edges. Social networks are dynamic i.e. there is continuous evolution and expansion. A node in a social network usually has several attributes. There are small and large communities within the social graph. Old graph analysis tools are not designed to manage such large and complex social network graphs.

Facebook is a preferred social network by marketers, not only because of the sheer number of users represented but also because of its incredibly insightful analytics suite. It’s important to be able to analyze customers and their behavior on a micro level due to Facebook’s ever-changing algorithm, and the implications for our content and business. If we refuse to adapt our approach based on these insights, we’re doomed to obscurity on the news feed.

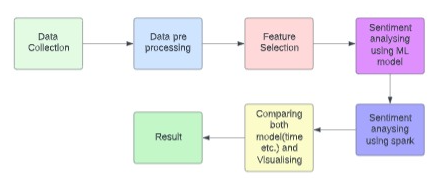
A deep Facebook data analysis shouldn’t be a one and done situation. Ideally, we’ll be auditing our efforts every few months or so at most. This will help us predict the likings, and the general summary of many users as a whole.

**Apache Spark:**

Apache Spark is a lightning-fast cluster computing system. It provides a set of high-level API namely Java, Scala, Python, and R for application development. Apache Spark is a tool for speedily executing Spark Applications. Spark utilizes Hadoop in two different ways – one is for Storage and second is for Process handling. Just because Spark has its own Cluster Management, so it utilizes Hadoop for Storage objectives.

Spark is intended to cover an extensive variety of remaining loads, for example, cluster applications, iterative calculations, intuitive questions, and streaming. Aside from supporting all these remaining tasks at hand in a particular framework, it decreases the administration weight of keeping up isolated apparatuses.

**Proposed system:**

****

**Explanation of project components:**

**Module 1: Data Collection**

Facebook Dataset is taken from the GitHub

* GitHub gives insights from Facebook dataset which consists of identifying users that can be focused more to increase the business.
* These valuable insights should help Facebook to make intelligent decisions to identify its useful users and provide correct recommendations to them. We have taken this dataset from github.com
* Here we used a csv file. This dataset contains 9903 entries (big dataset) with 14 columns. We have columns namely age, userid, dob, etc., Column names are well defined so that everyone can interpret easily

<https://github.com/jaegoan9/Facebook-Sentiment-Analysis>

**Module 2: Data Pre-Processing**

The collected data consists of different emotions, stop words, acronyms, etc. But during analysis this type of data needs to be converted into the proper format to extract sentiments from the user behavior.

* Tokenization
* Various Dictionaries

1. Acronym Dictionary
2. Stop Words Dictionary

* Emoticon

**For Example:**

Considering one of the Facebook posts regarding new mobile features. Users' opinion about the new phone might be positive or negative or neutral.

**Example for Positive Sentiment**

Looks are awesome. Battery backup is excellent. Camera is good. The display light quality is good.

**Example for Neutral Sentiment**

Although this is good mobile, looks good, but Problem is that it doesn’t provide separate Space for dual SIM & memory card together.

**Example for Negative Sentiment**

Not good one as expected. Camera quality very poor.

1. **Tokenization**

Comments extracted from Facebook are divided into tokens. This is known as tokenization process.

For example, ‘Looks are awesome. Battery backup is excellent. Camera is good. The display light quality is good is divided down into ‘Looks’, ‘are’, ‘awesome’, ‘.’, ‘Battery’, ‘backup’, ‘is’, ‘excellent’, ‘.’, ‘Camera’, ‘is’, ‘good’, ‘.’, ‘The’, ‘display’, ‘light’, ‘quality’, ‘is’, ‘good’, ‘.’

1. **Various Dictionaries:**

**Acronym Dictionary:**

It is used to give the required acronym for the words, if needed.

**Stop Words Dictionary:**

It is used to remove the unrelated words in the sentiment analysis process. Example: A, An, The, Has, Are, Is.

1. **Emoticon:**

This is used to detect the emoticons for the purpose of classifying the comment as positive or negative or neutral.

**Module 3: Sentiment Analysis**

The user sentiments collected from the Facebook are categorized into positive, negative, neutral. This sentiment analysis can be performed for different purposes based on the business objectives.

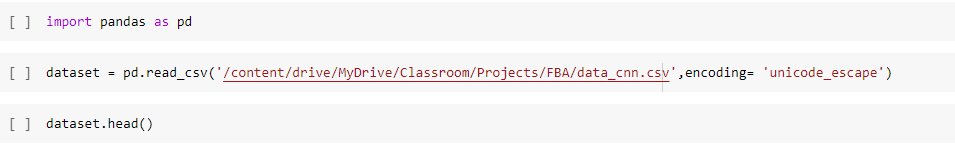
**Module 4: Data Visualization**

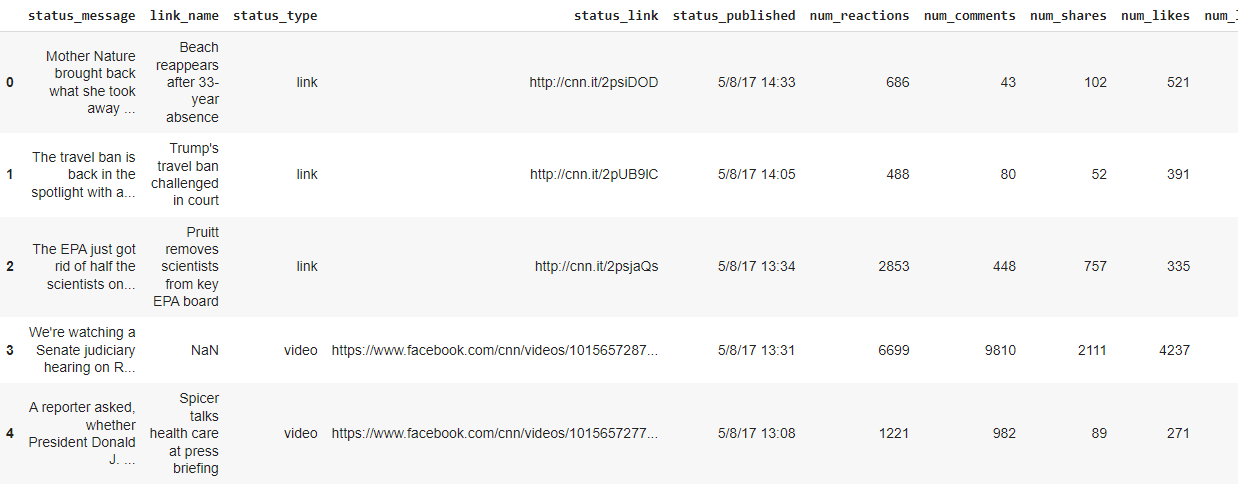
After the Facebook sentiment analysis, the extracted and analyzed sentiments are visualized using Tableau.

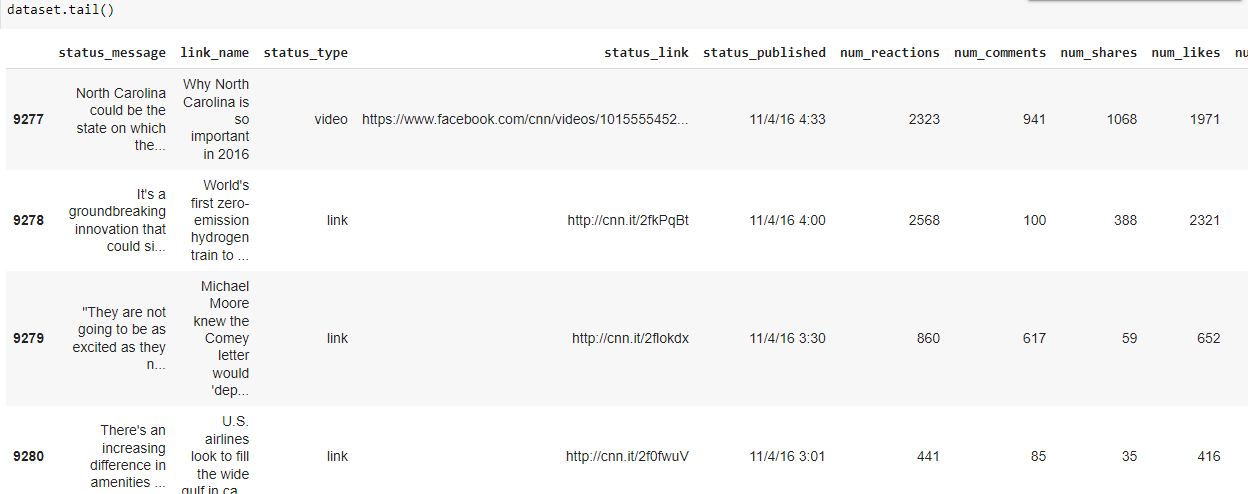
**Deployment:**

**Sentimental Analysis using Spark**

**Load data**

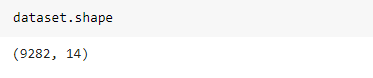
****

****

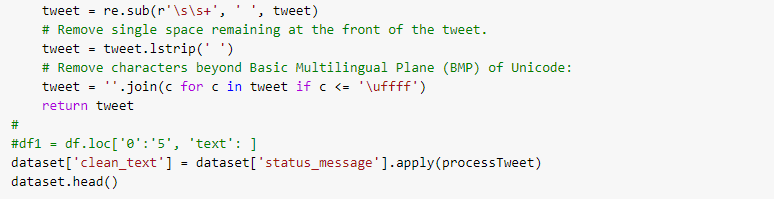
****

****

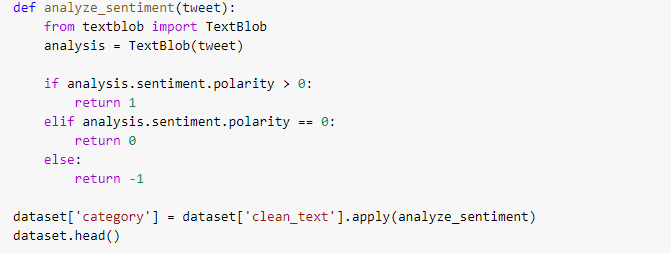
**Preprocessing**

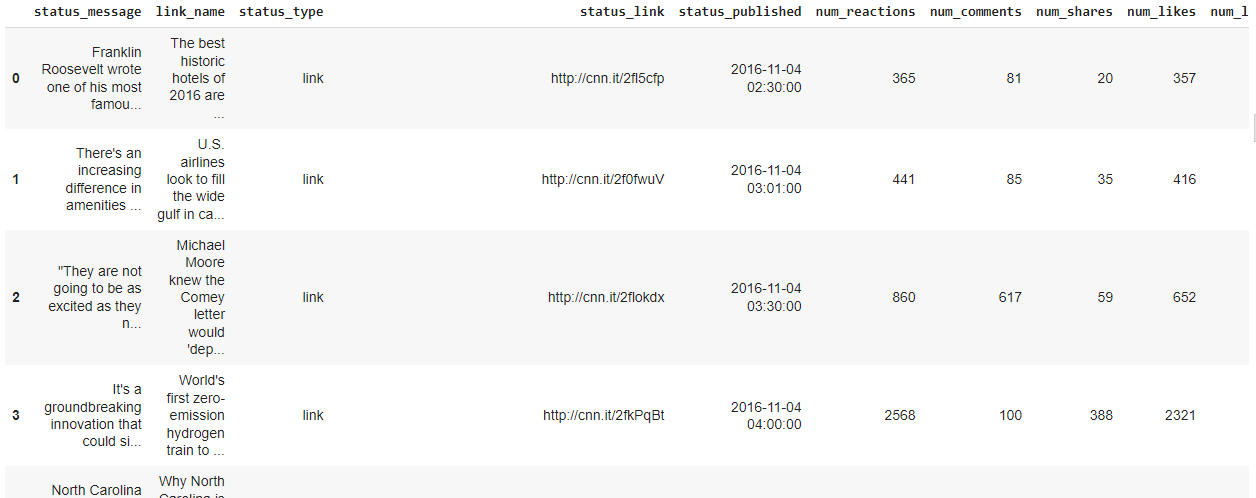
****

****

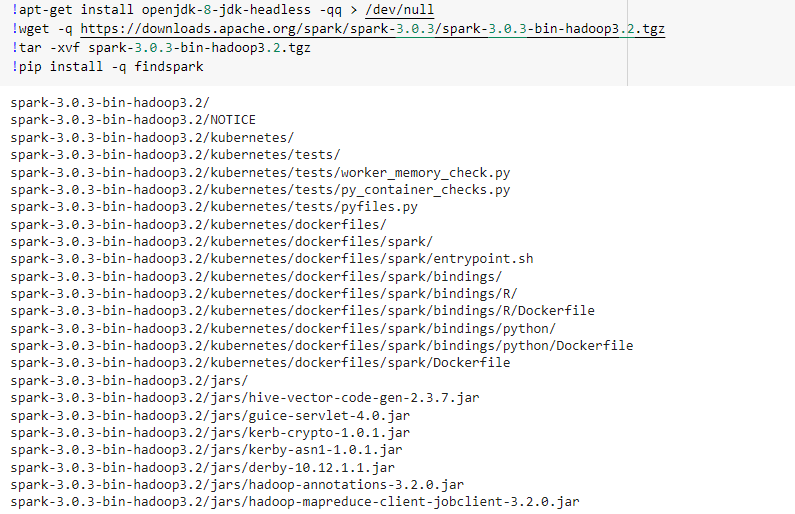
****

****

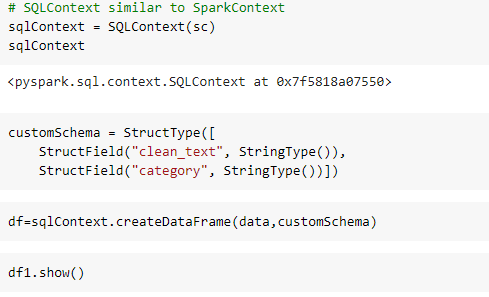
****

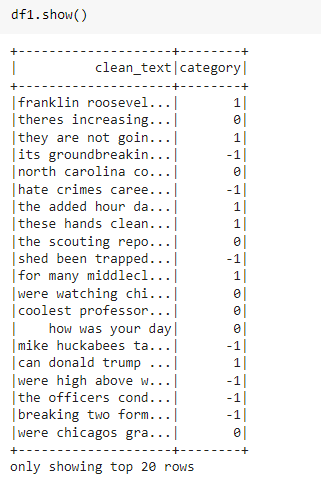
****

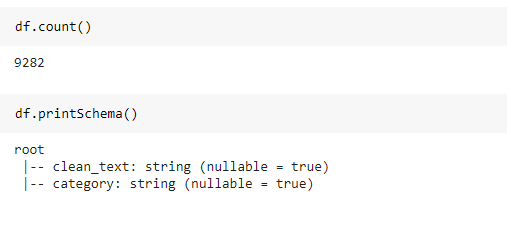
****

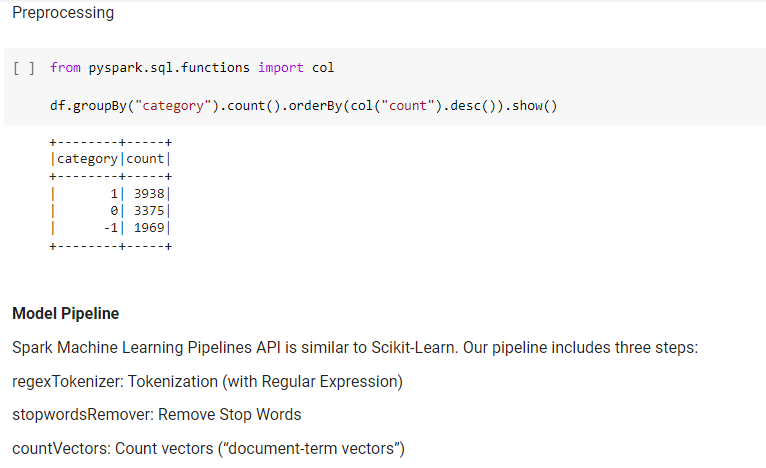
****

****

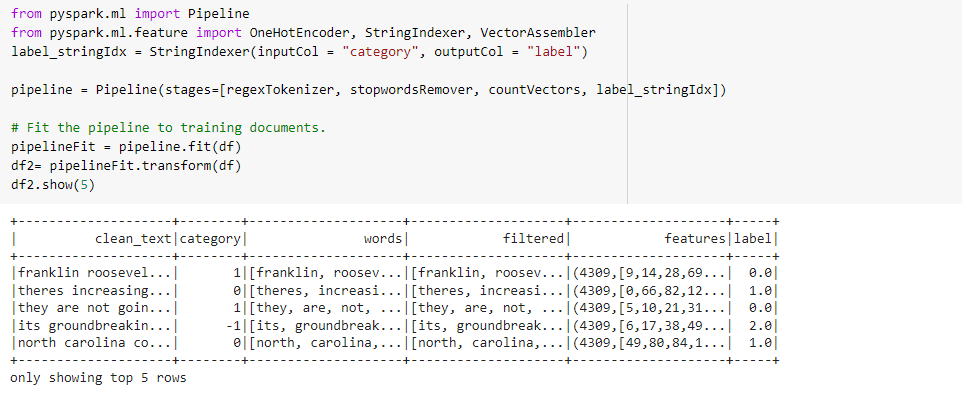
****

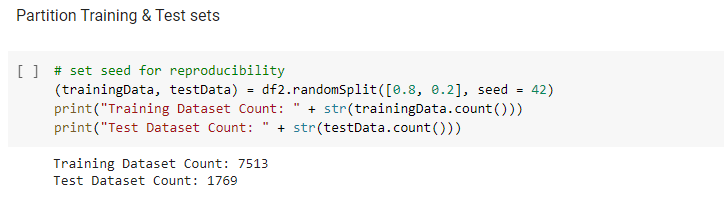
****

****

****

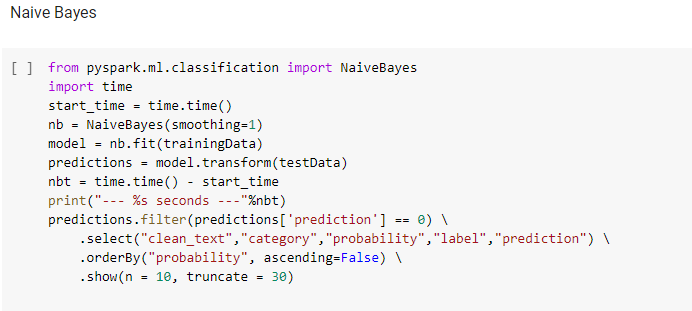
****

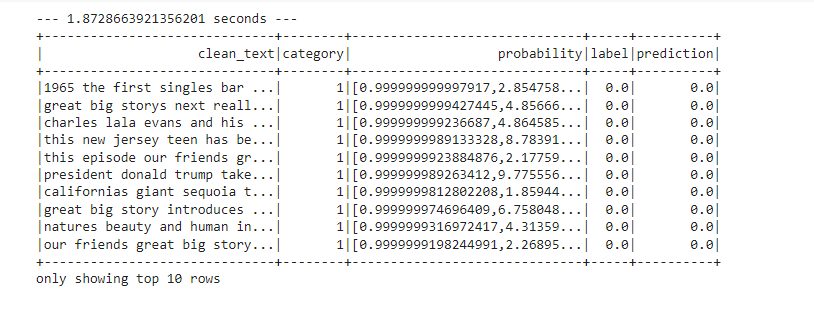
****

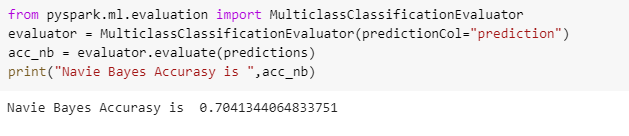
****

**ML Algorithm using pyspark**

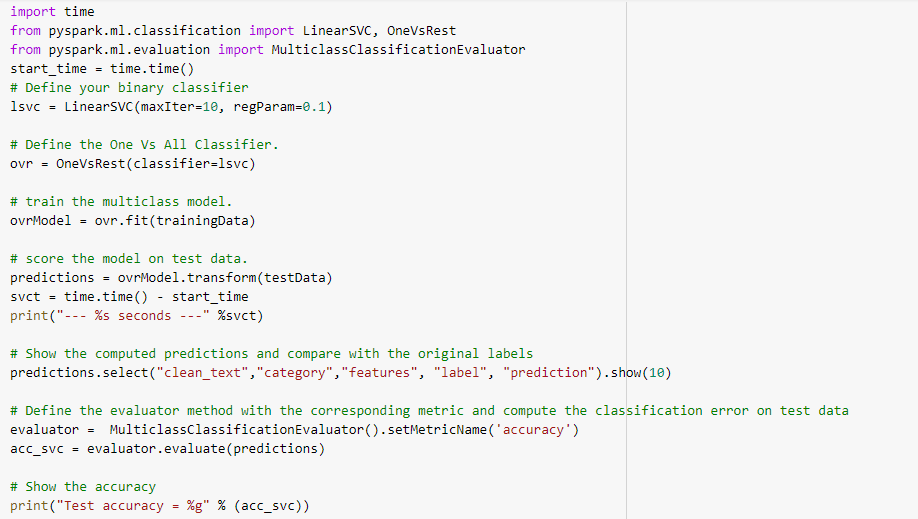
**Naïve bayes**

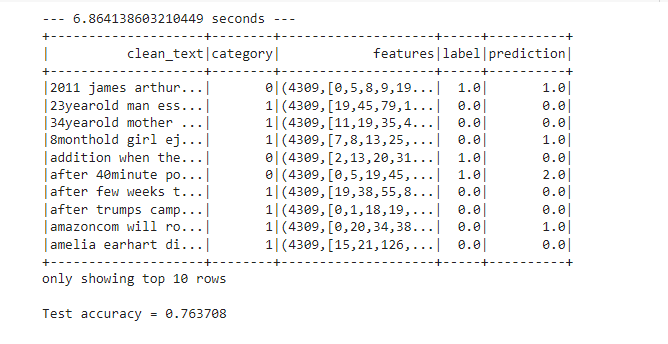
****

****

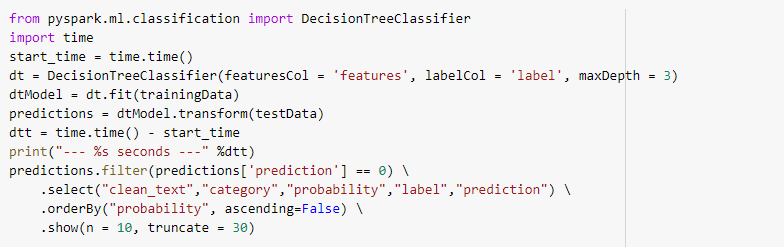
****

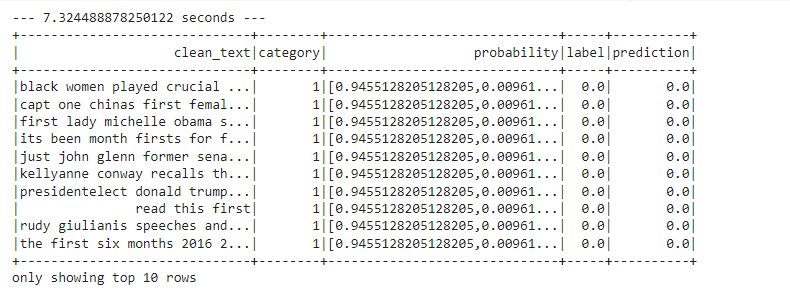
**SVC**

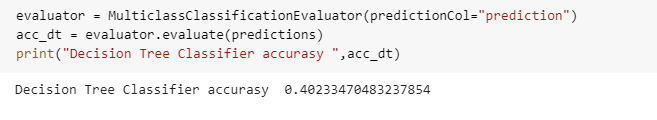
****

****

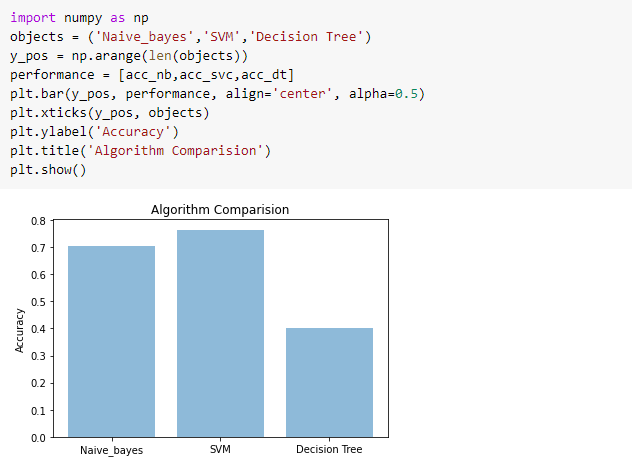
**Decision Tree**

****

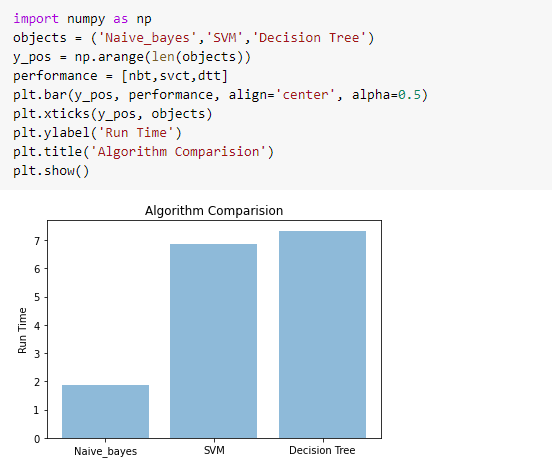
****

****

**Accuracy comparison**

****

**Time Comparisons**

****

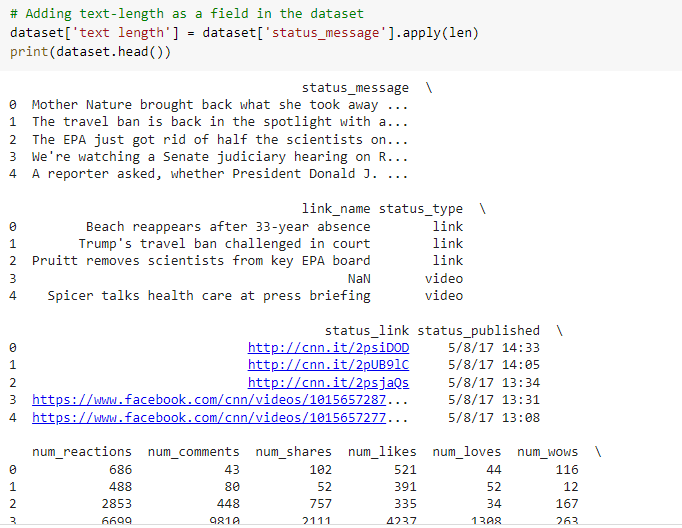
**Sentimental Analysis using Python**

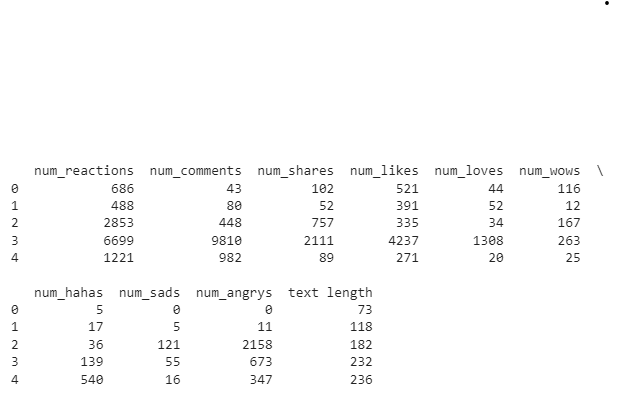
**Load data**

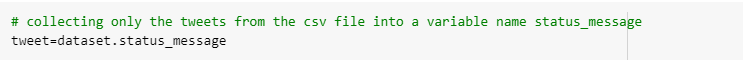
****

**Preprocessing**

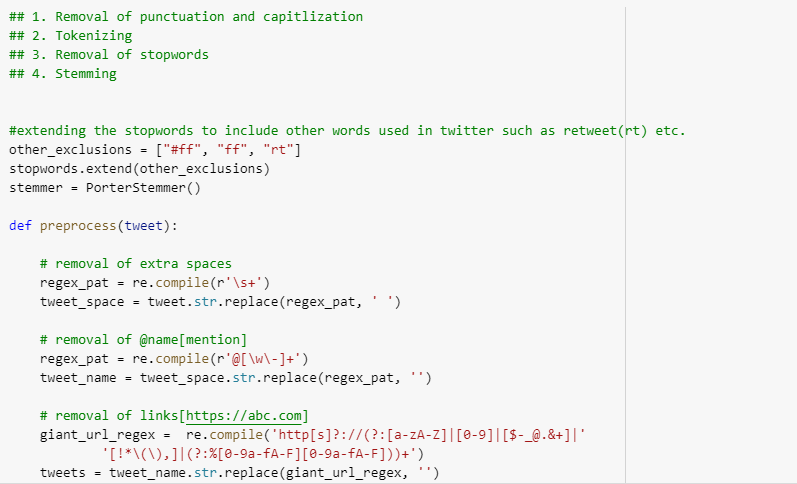
****

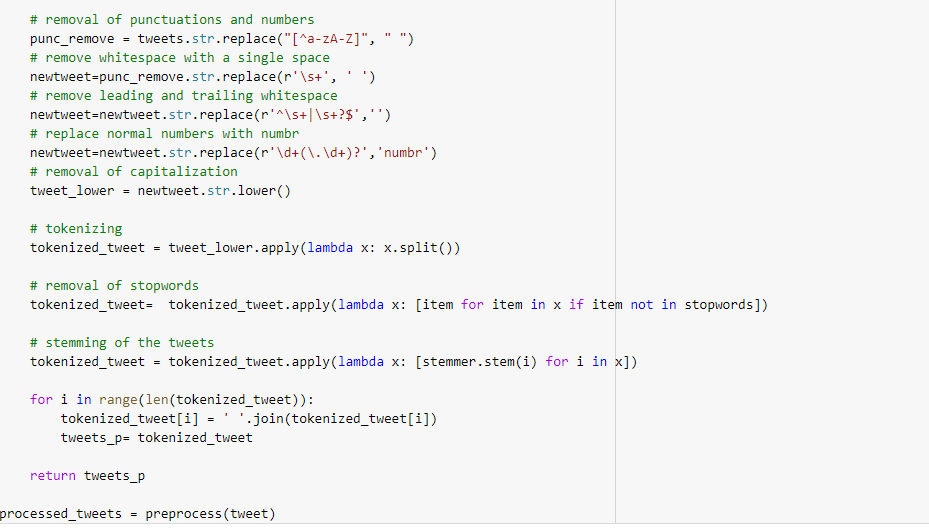
****

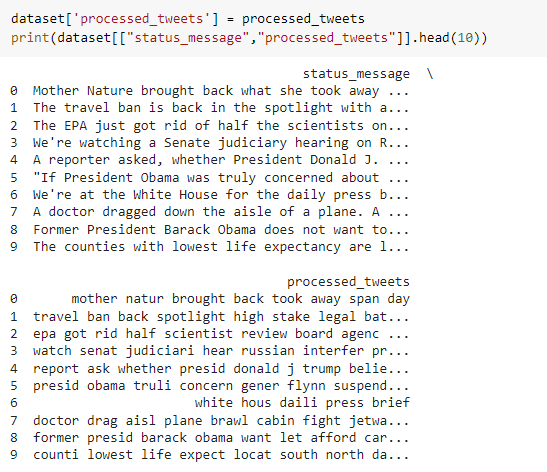
****

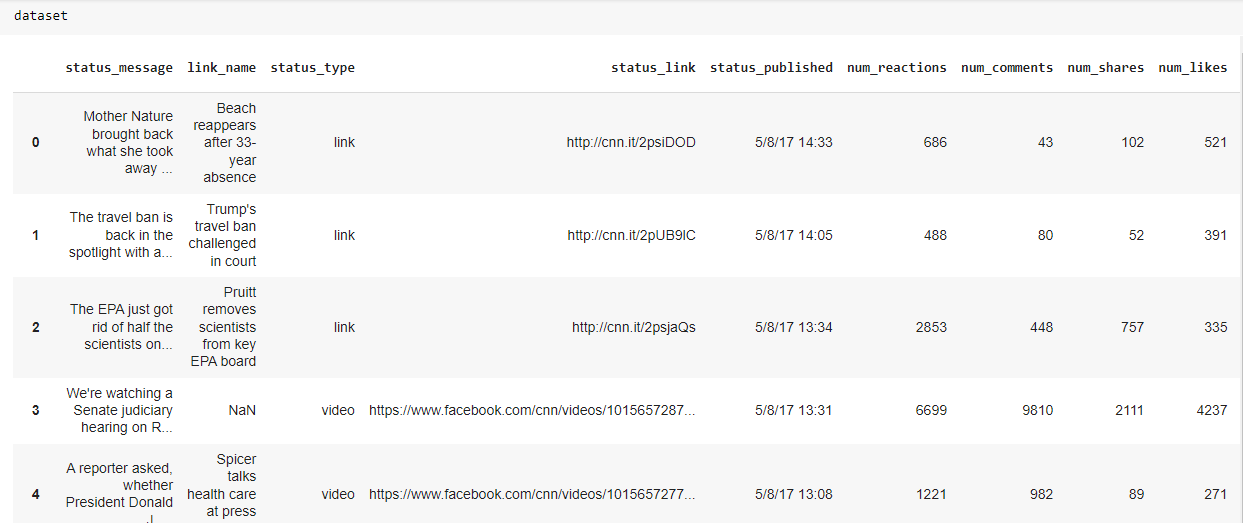
****

****

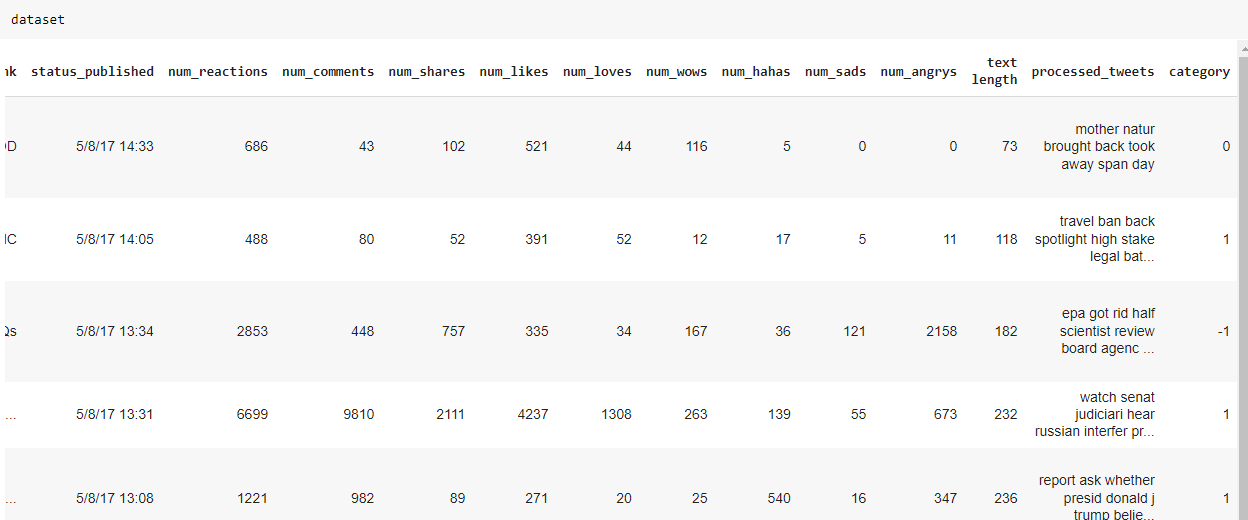
****

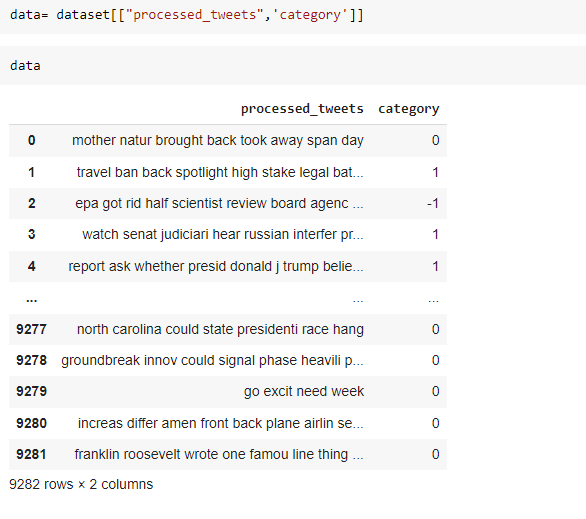
****

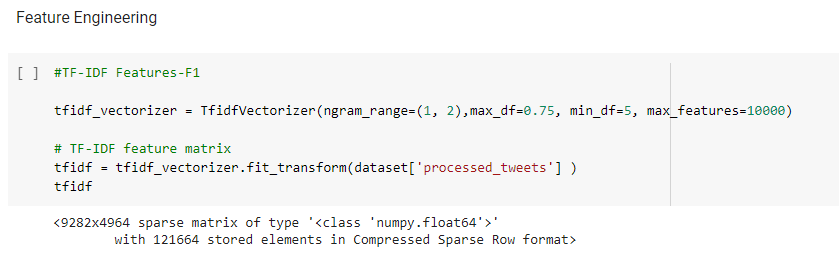
****

****

****

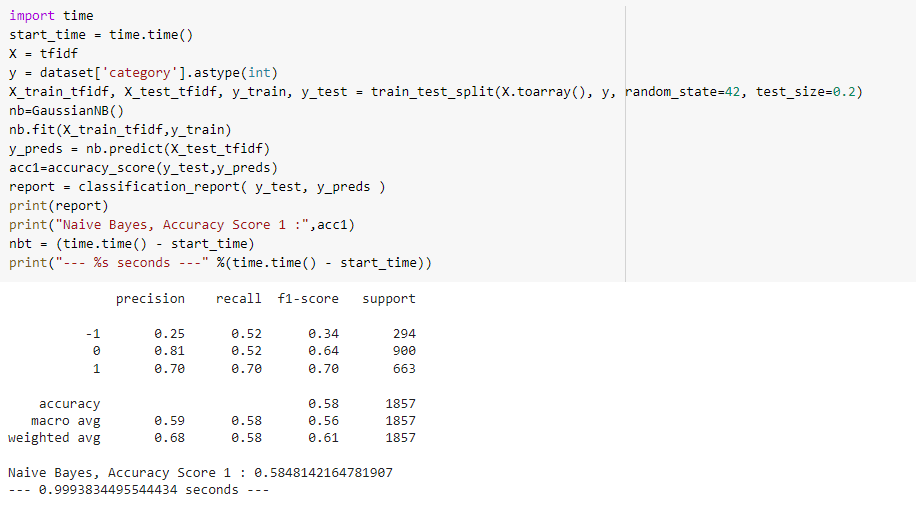
****

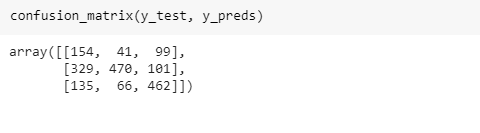
****

****

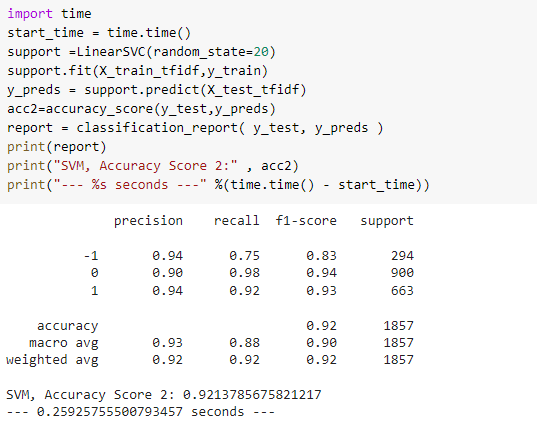
**ML Algorithm**

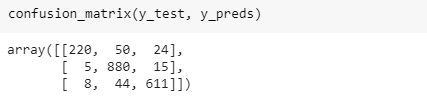
**Naïve bayes**

****

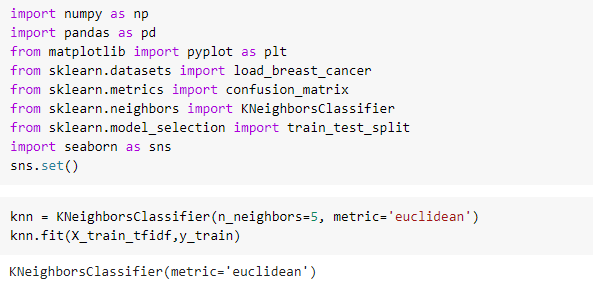
****

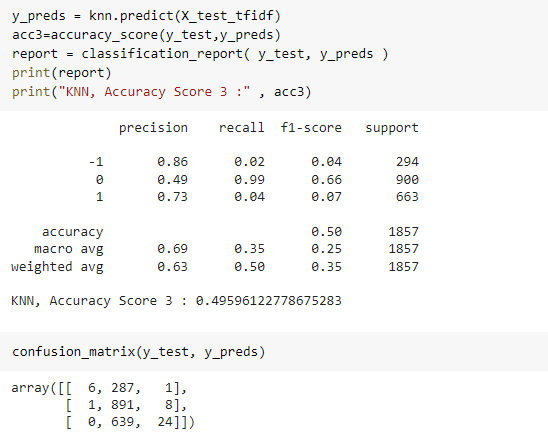
**SVM**

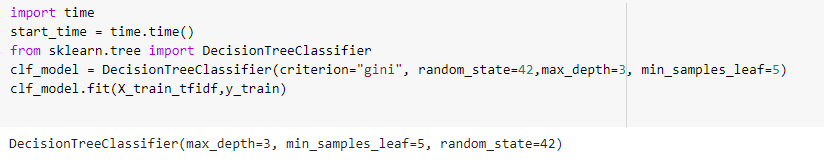
****

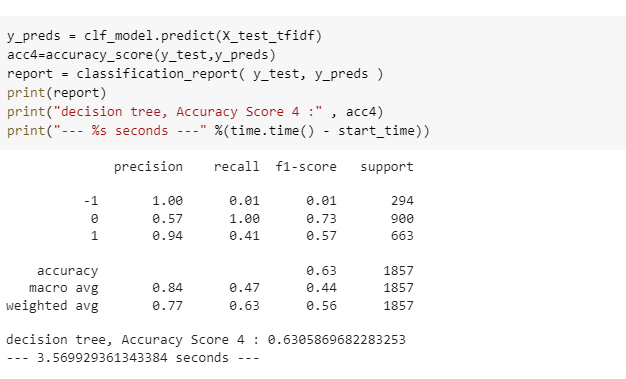
****

**KNN**

****

**Decision Tree**

****

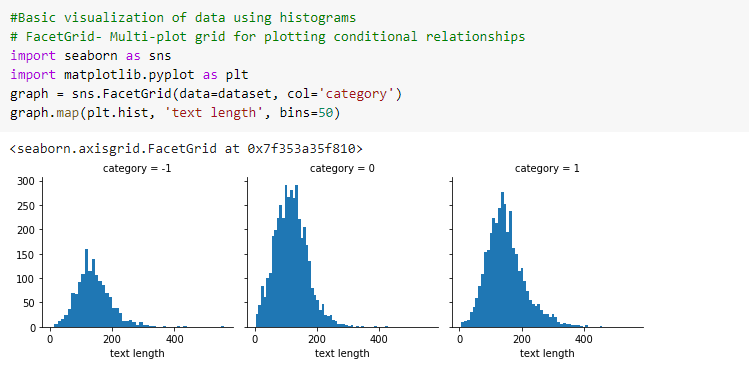
****

**Algorithm comparison**

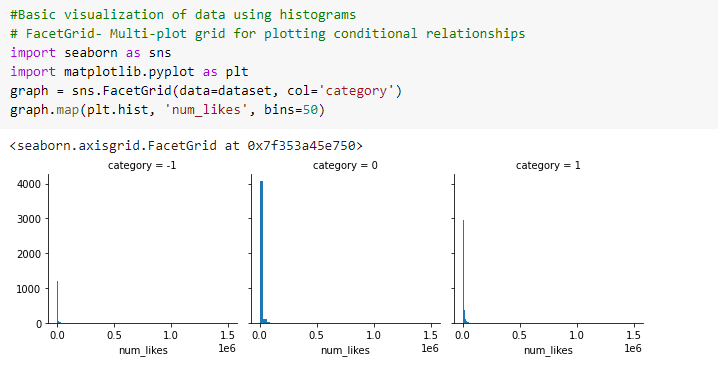
****

**Visualization**

**Histogram**

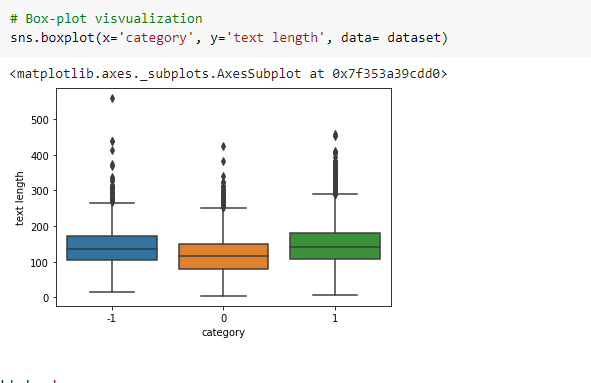
****

Histogram between text length and category , there is no category distributed normally except category 1 . category 1 distributed normally.

****

This Histogram shows relationship between category and number of likes .

**Boxplot**

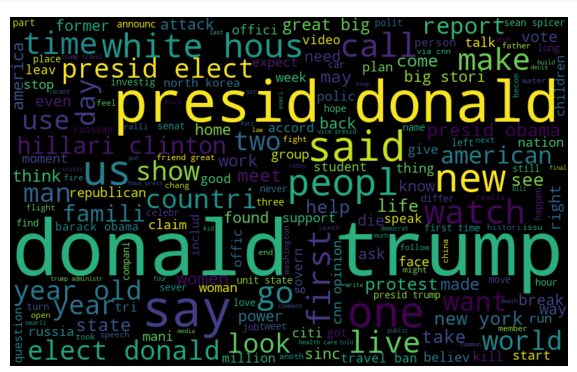
****

This shows boxplot between category and text length. Negative comment have more text length.

**Word Cloud**

**Common words used in dataset**

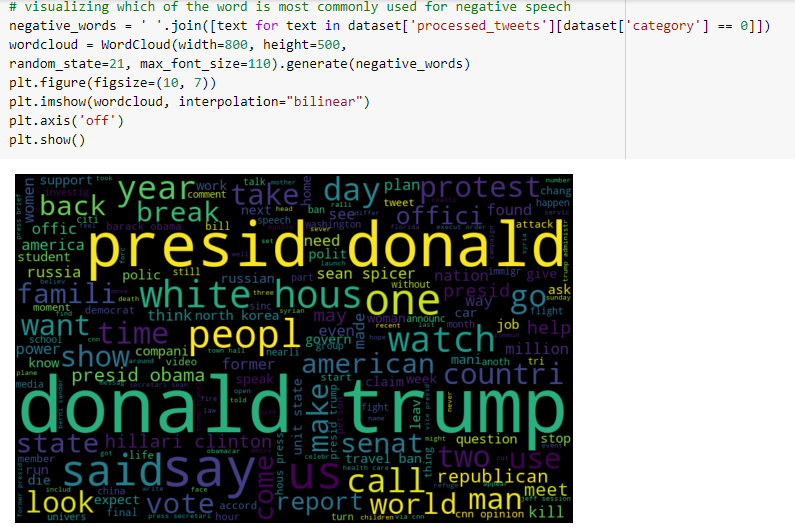
****

****

**Words used in positive text**

****

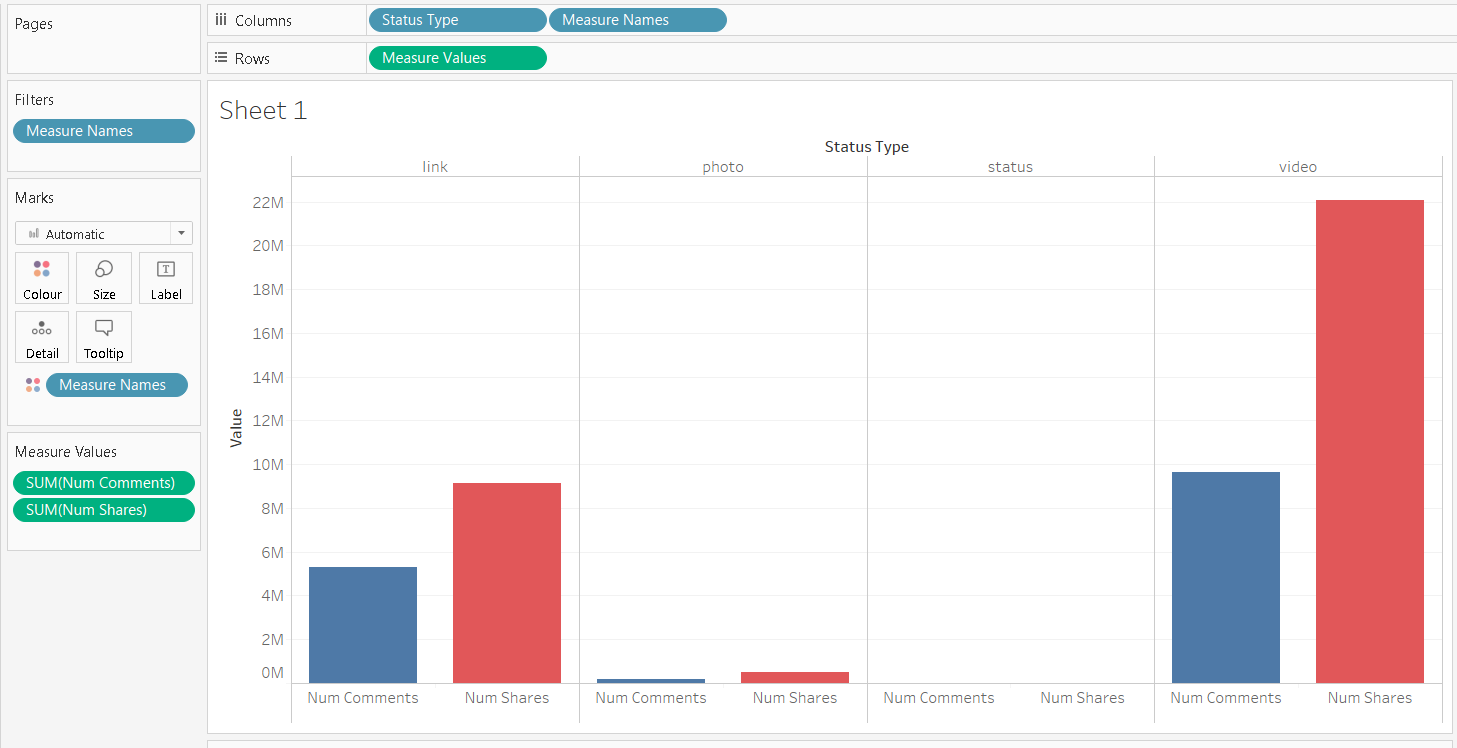
**Words used in negative commends**

****

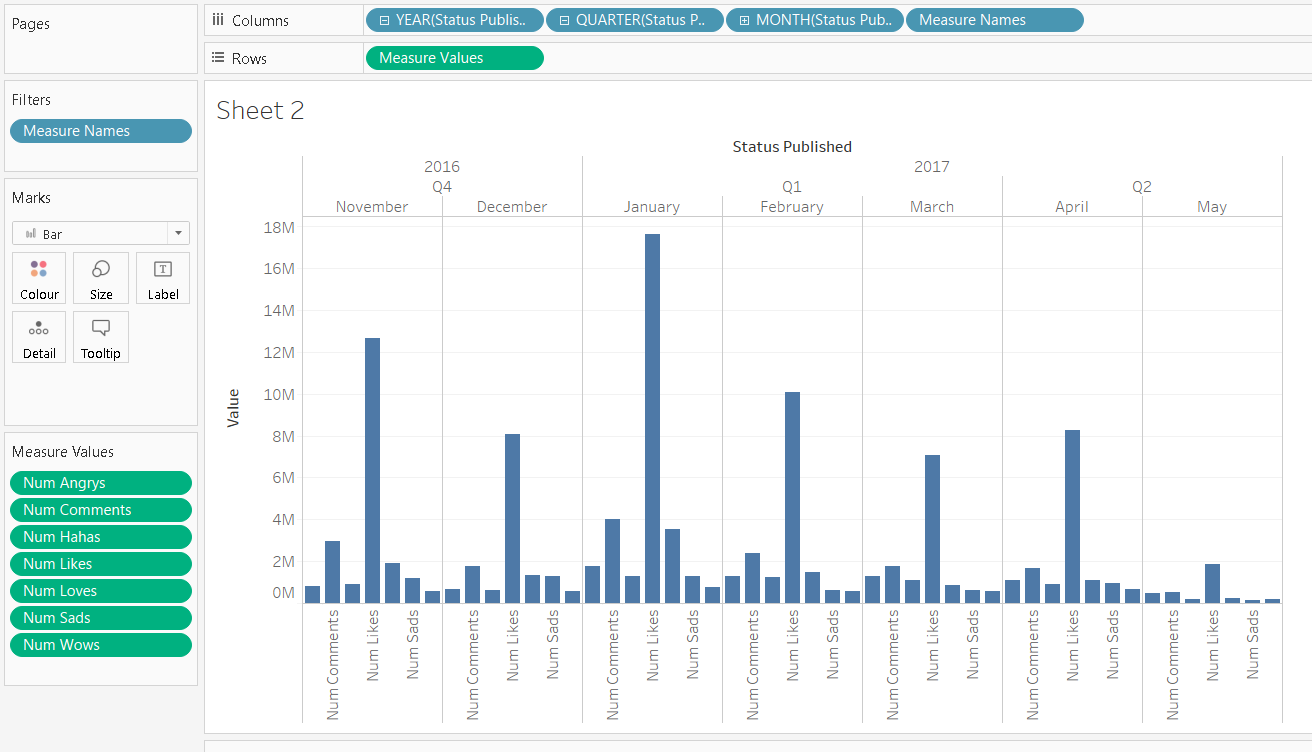
**Words used in neutral comment**

****

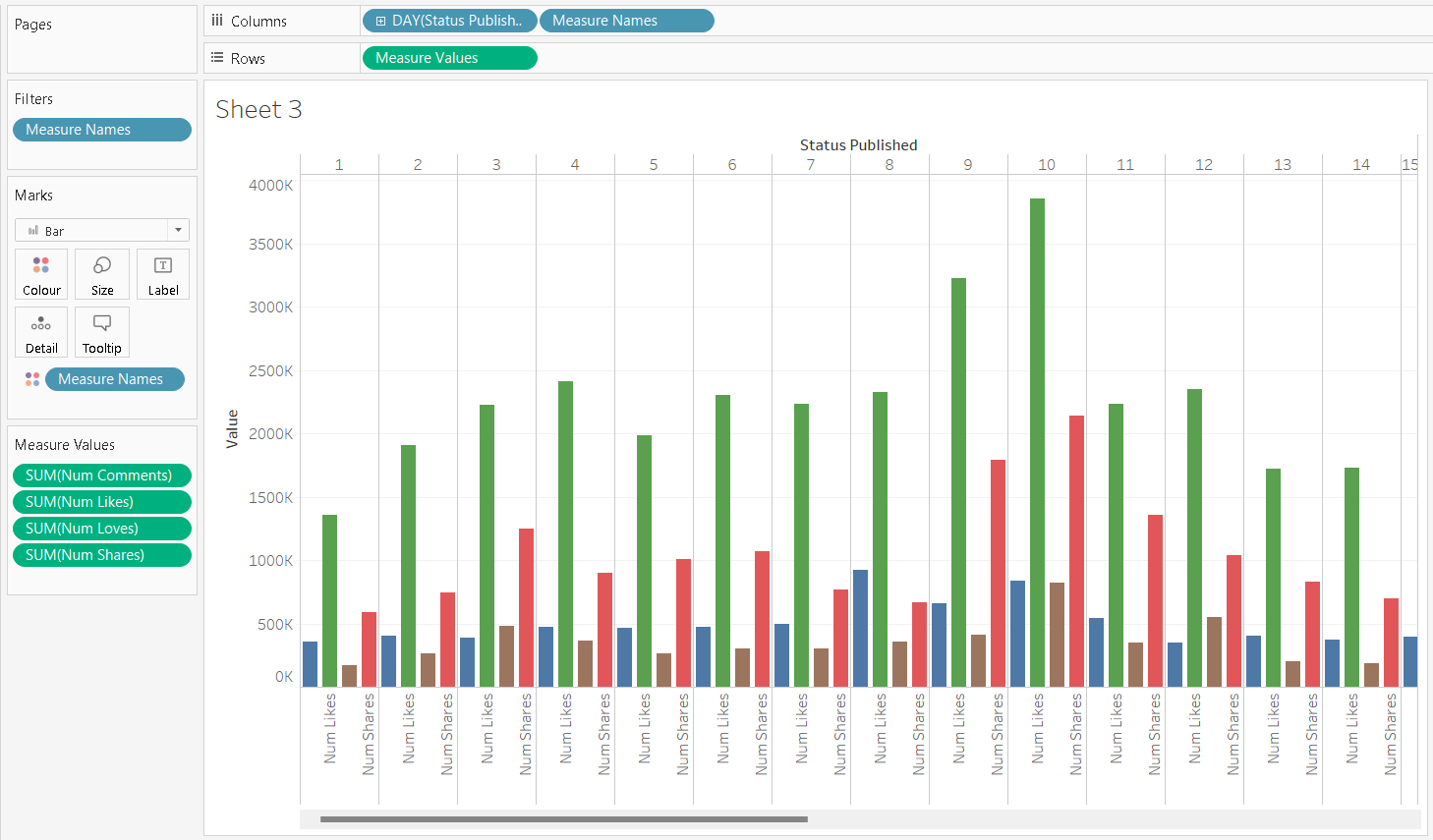
**Visualization using Tableau**

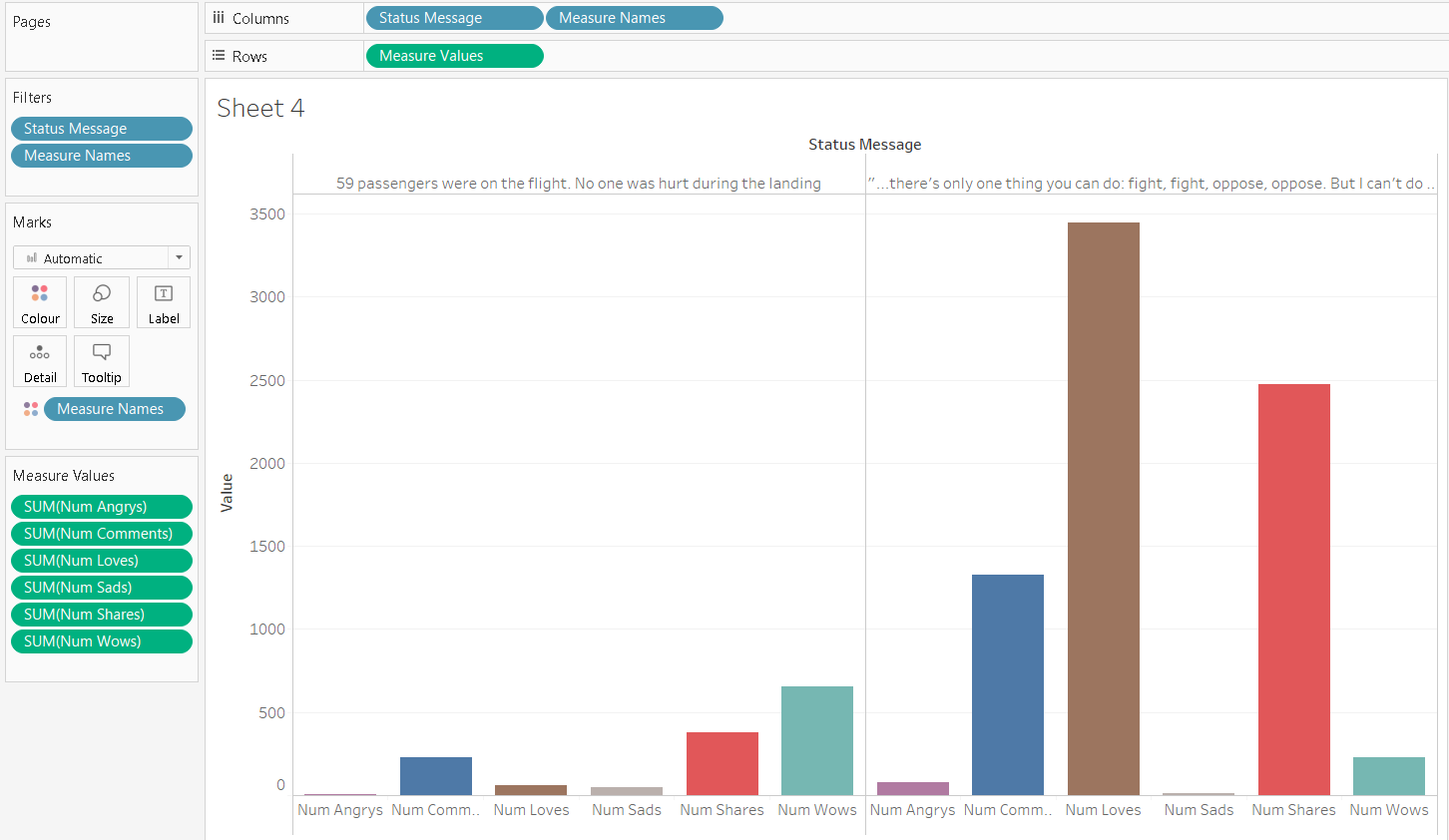
****

**From the above graph we conclude that num\_shares, num\_comments is more for the videos compare to link, photo, status. If we want to do any adds or to attract the audience then we can use the video format.**

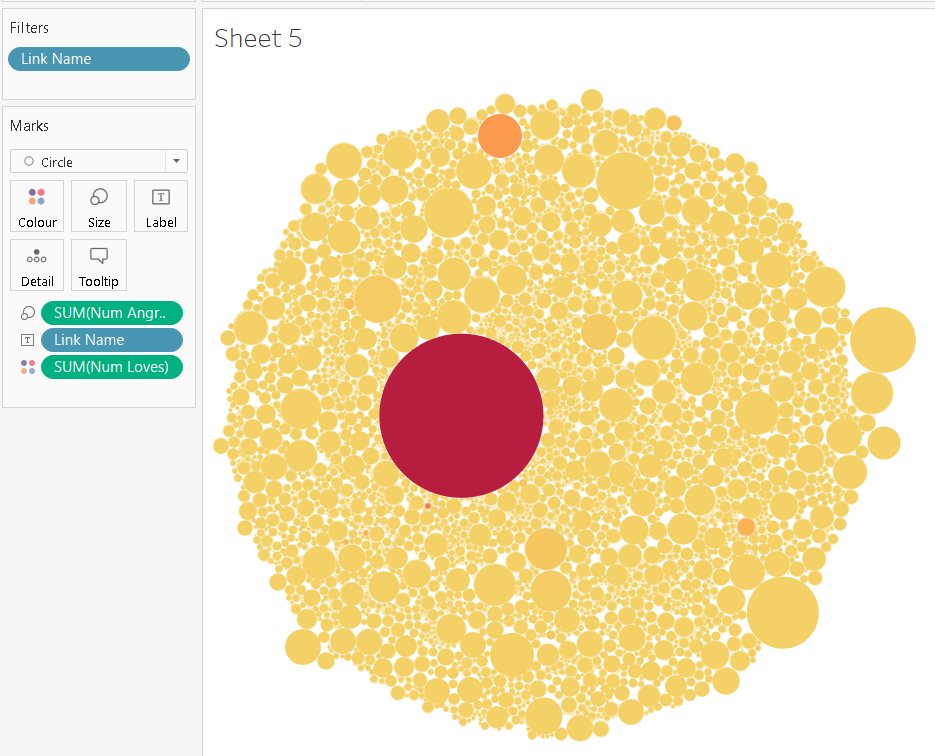
****

**From the above graph we can say that Dec-Jan num\_likes is very high. Nov - Dec is the months of New year and Christmas so there is many intersing posts,sales,etc..**

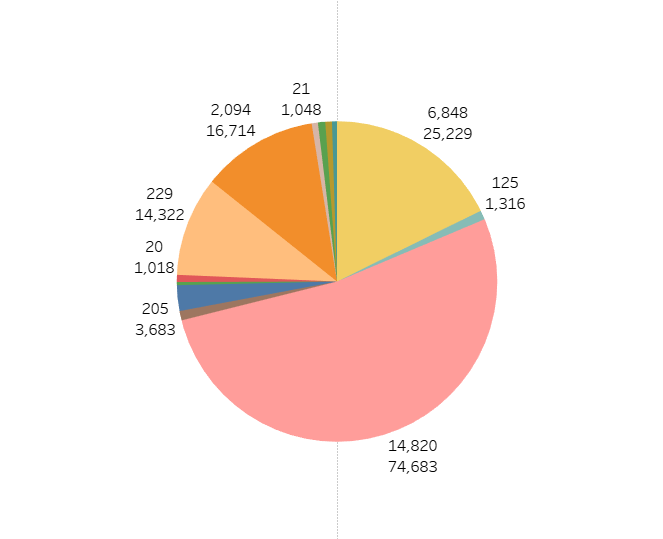
****

****

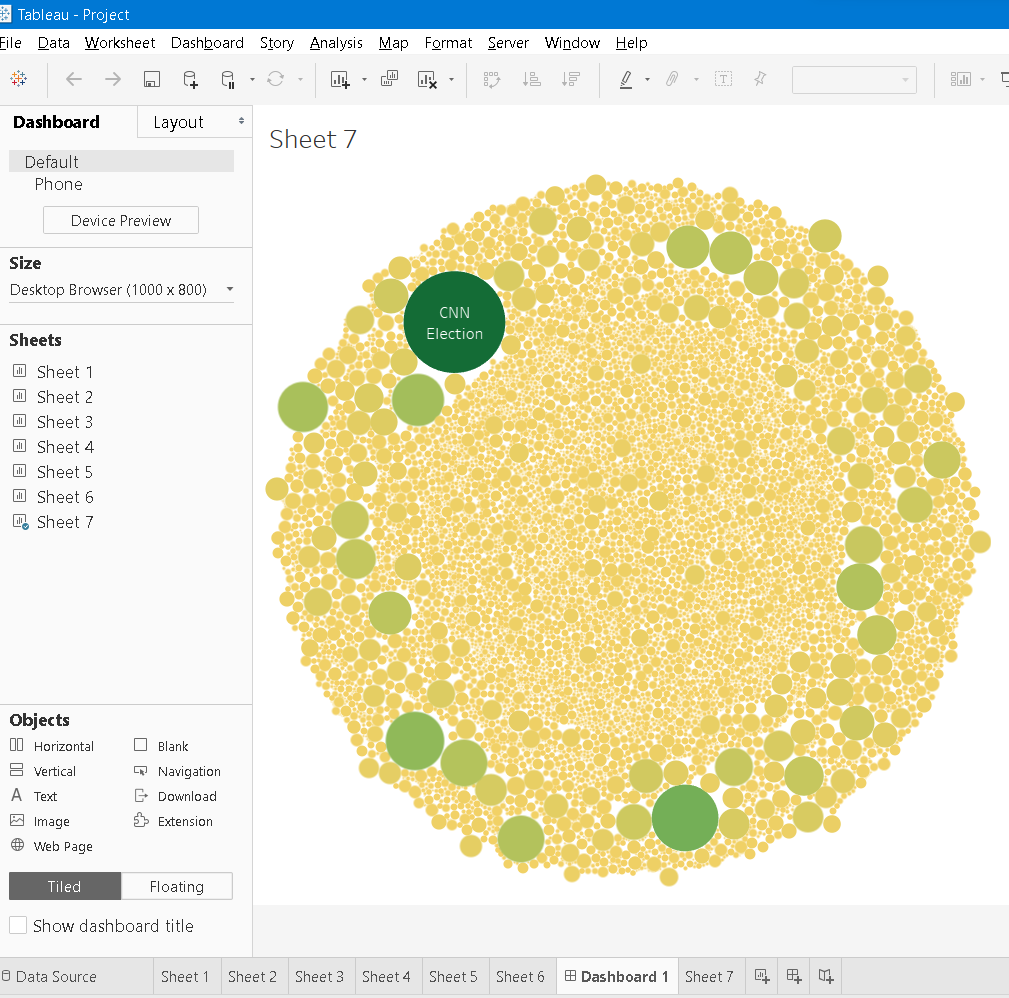
**Bubble Chart:**

****

**Pie Chart**

****

**Dashboard:**

****

**In the above Dashboard it is directly connected to web in such a way When we click on some point then it will directly go to chrome and it open that point related page. In this way**

****

**DNN:**

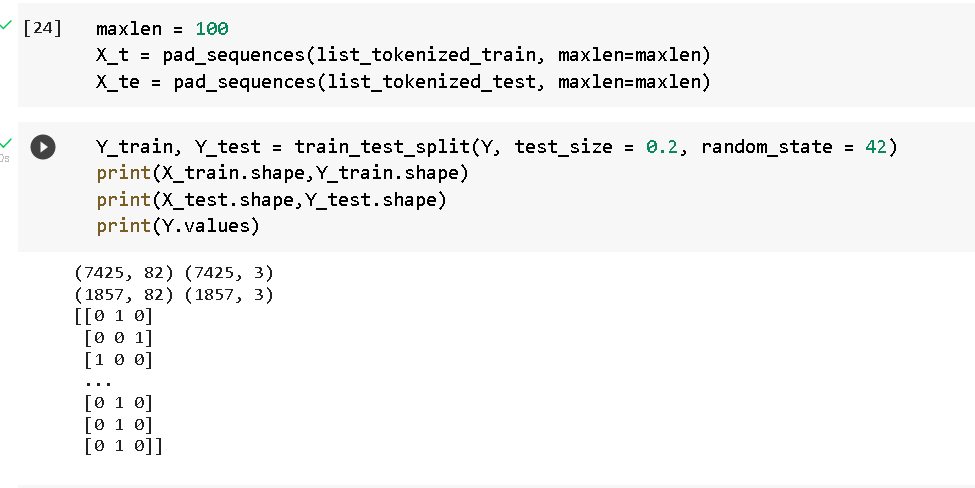
After preprocessing split the data into x and y

****

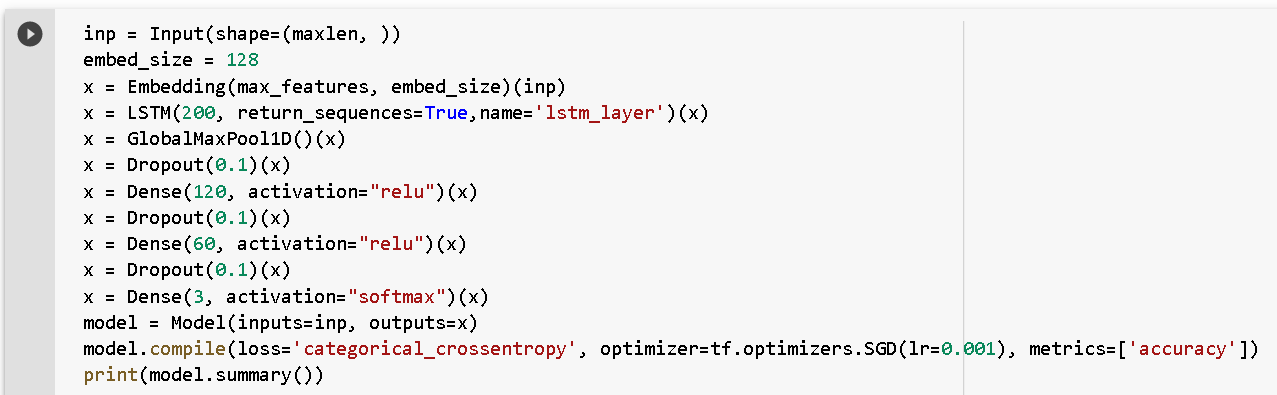
After that split the x and y dataset into train and test

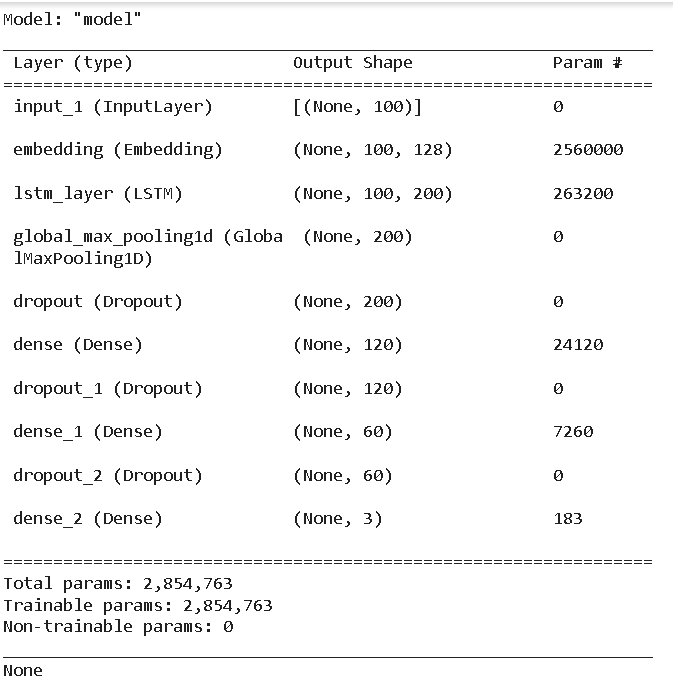
****

Convert the Xtrain and Xtest to sequence

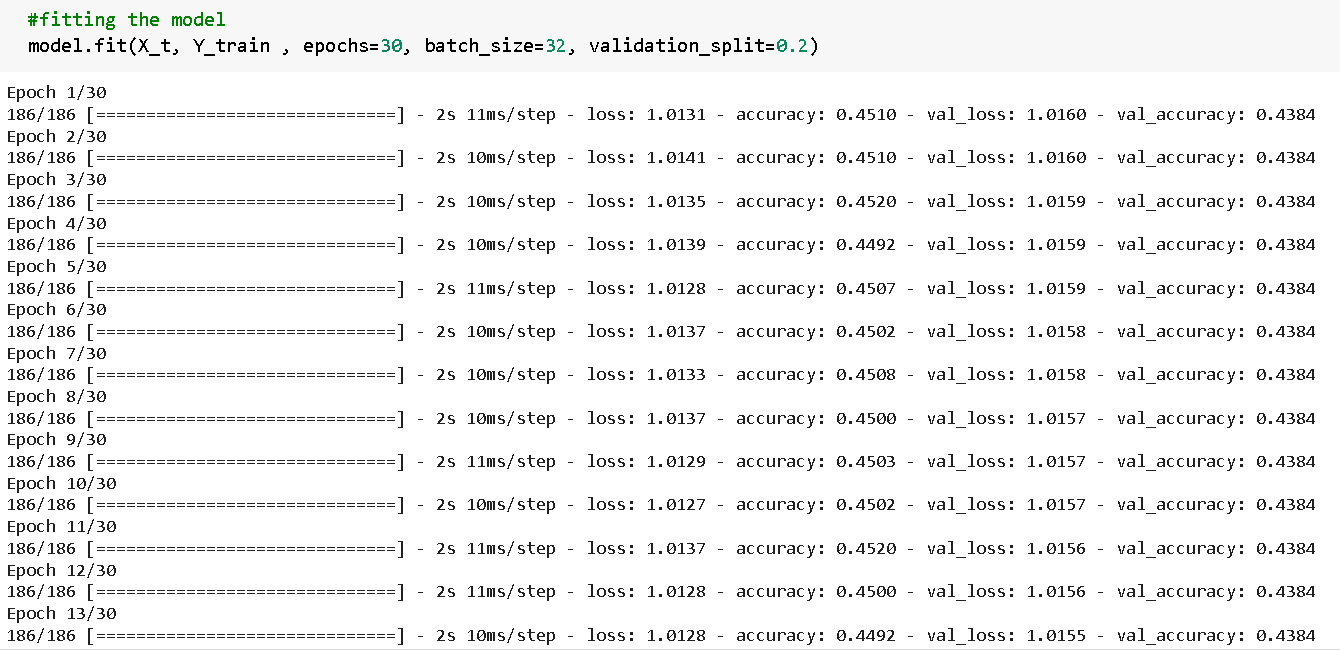
****

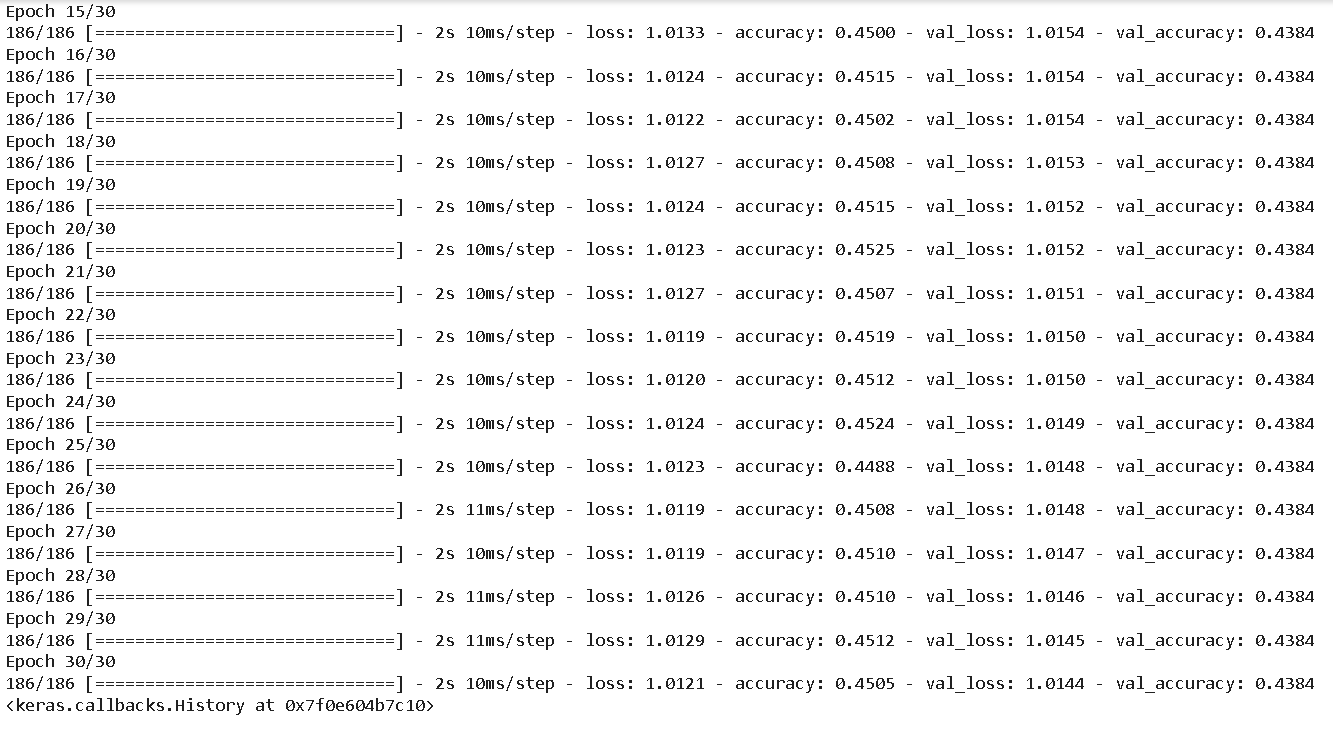
Creating the model:

****

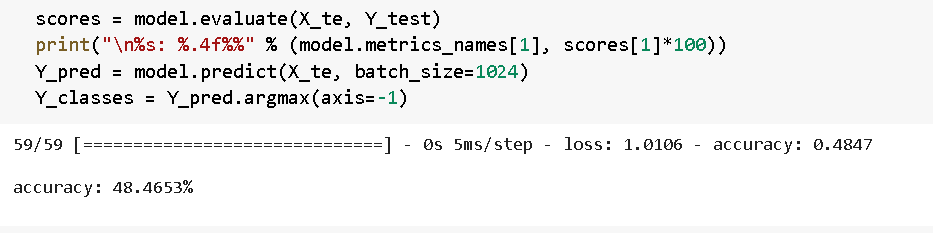
****

Next is the Fitting the model

****

****

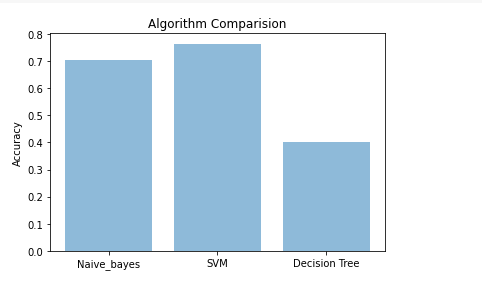
Next is the evaluating the model:

****

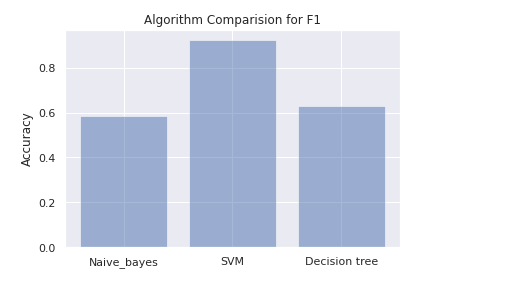
**Results:**

Accuracy comparison between pyspark ML algorithm and ML algorithm.

* ML using pyspark



* ML algorithm



Accuracy comparison between pyspark ML algorithm and ML algorithm.

* ML using pyspark



* ML algorithm
* Naive Bayes, Run time 0.9993834495544434
* SVM, Run time 0.25925755500793457
* decision tree, Run time 3.569929361343384

**Conclusion and future work:**

* We compared Accuracy and Run time of normal ML algorithm and ML algorithm using pyspark.
* SVM got high accuracy in both methods. So SVM is the best model for this dataset.
* Comparing normal and pyspark ML model there is no much difference.
* Both giving accuracy with point difference.
* When we compared Run time between two methods normal ML models took less time.
* Because our dataset size is 10,000. It will different when we using large dataset.
* From this we can know spark will give good accuracy with less runtime when we using large dataset.
* In future we will try to use different Deep learning models with pyspark

**References:**

* [**https://towardsdatascience.com/an-easy-tutorial-about-sentiment-analysis-with-deep-learning-and-keras-2bf52b9cba91**](https://towardsdatascience.com/an-easy-tutorial-about-sentiment-analysis-with-deep-learning-and-keras-2bf52b9cba91)
* [**https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python**](https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python)
* [**https://www.analyticsvidhya.com/blog/2021/08/text-preprocessing-techniques-for-performing-sentiment-analysis/**](https://www.analyticsvidhya.com/blog/2021/08/text-preprocessing-techniques-for-performing-sentiment-analysis/)
* [**https://www.section.io/engineering-education/multiclass-text-classification-with-pyspark/**](https://www.section.io/engineering-education/multiclass-text-classification-with-pyspark/)
* [**https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa**](https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa)