**Experiment No: 3**

**Aim:** Data Cleaning and Storage- Preprocess, filter and store social media data for business

**Objective:** To perform preprocessing on social media data and make it ready for analysis.

###### Lab outcomes:

*At the end of this lab session, students will be able to…*

1. Clean and preprocess the data captured from social media.
2. Perform the exploratory data analysis.

**Theory:**

● Data cleaning and preprocessing is an essential – and often crucial – part of any analytical process. Social media contains different types of data: information about user profiles, statistics

● (number of likes or number of followers), verbatims, and other media content.

● Quantitative data is very convenient for an analysis using statistical and numerical methods, but unstructured data such as user comments is much more challenging.

● To get meaningful information, one has to perform the whole process of information retrieval. It starts with the definition of the data type and data structure.

● On social media, unstructured data is related to text, images, videos, and sound and we will mostly deal with textual data.

● Then, the data has to be cleaned and normalized.

**Preprocessing**

● Preprocessing is one of the most important parts of the analysis process.

● It reformats the unstructured data into uniform, standardized form.

● The characters, words, and sentences identified at this stage are the fundamental units passed to all further processing stages.

● The quality of the preprocessing has a big impact of the final result on the whole process.

● There are several stages of the process: from simple text cleaning by removing white spaces, punctuation, HTML tags and special characters up to more sophisticated normalization techniques such as tokenization, stemming or lemmatization.

**Steps :**

Step 1: Loading Packages

Step 2: X’s Data extraction using snscrape Python library

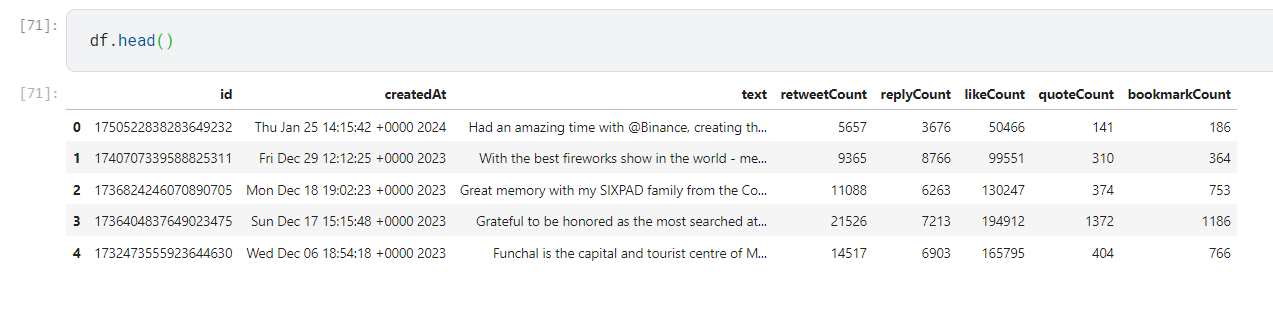
Step 3: X’s Data Cleaning and Preprocessing using Python

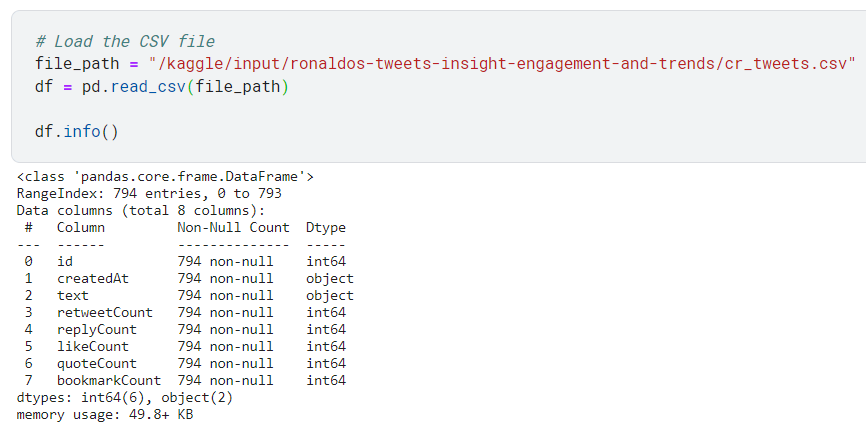
Step 4: X’s Data Visualization

Step 5: Twitter Data Sentiment Analysis using Textblob.

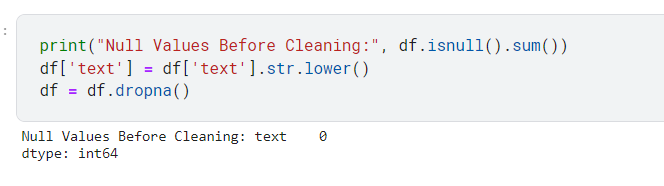
**Student’s Tasks**

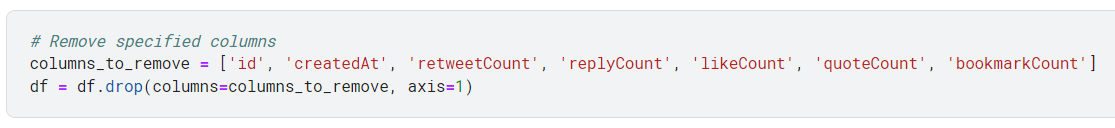
1. Scrape X’s Data for Ronaldos tweets insight engagement and trends 2024
2. Create document corpus with tweet text





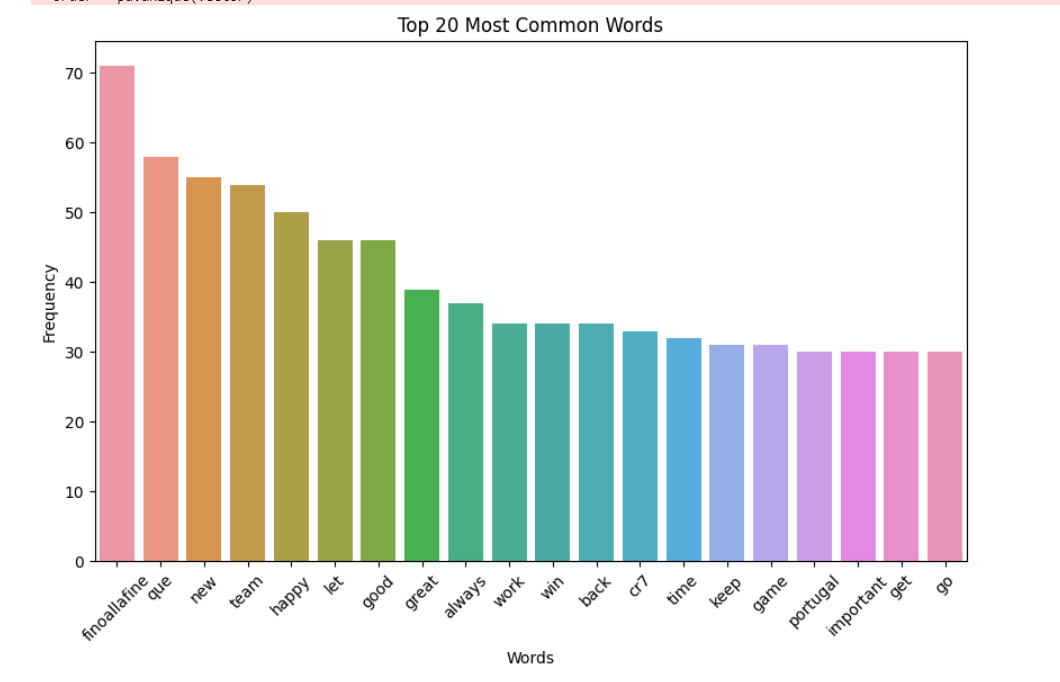
1. Data Cleaning & Preprocessing-
   1. Convert text to Lower Case
   2. Remove the links (URLs)
   3. Remove anything except the English language and space.
   4. Remove Stop words.







1. Visualize top 20 most common words.

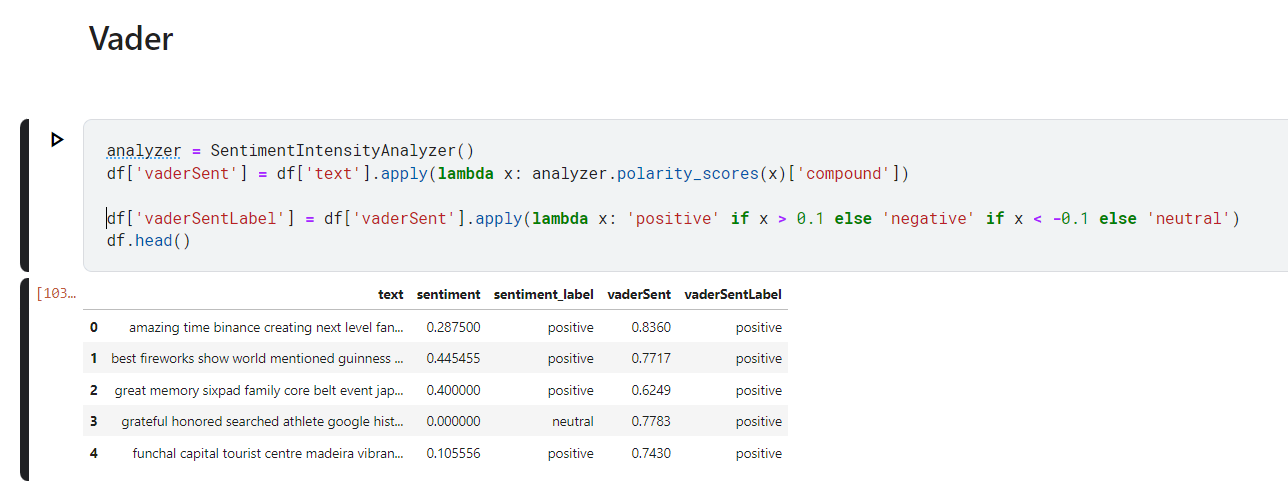


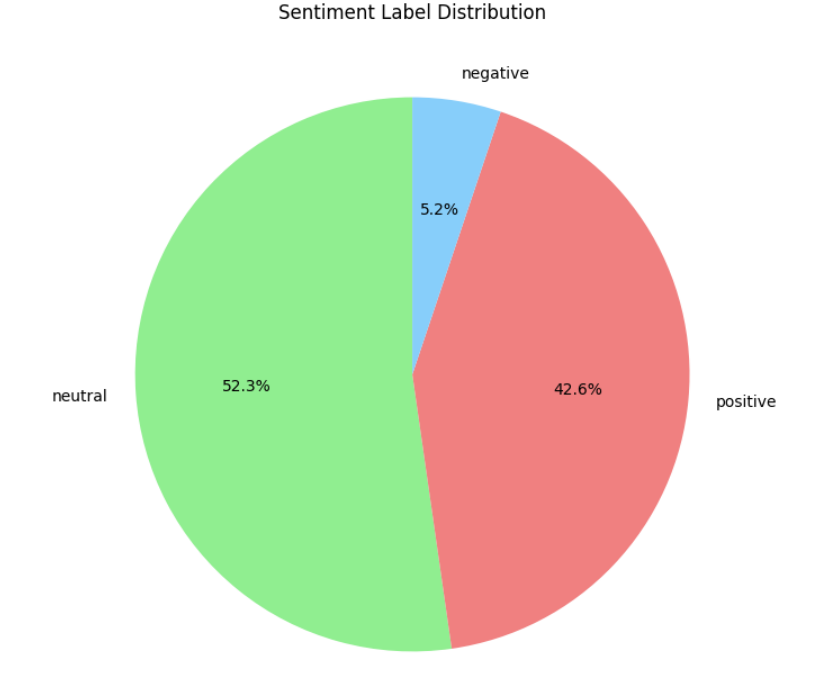
1. Visualize top 10 bigrams.



1. Perform the sentiment analysis using textblob.







1. Display the word cloud of positive words.



1. Display the word cloud of negative words.



1. Display the word cloud of neutral words.



**Kaggle Link (for Code) :**

<https://www.kaggle.com/mrappplg/sma-exp-3-v1>