

**Fundamentals of Big Data Analysis**

**Assignment# 3**

Part - 1

**French presidential election 2022**

**Muhammad Shahzad Waris**

**22L-7530**

**BDS-4B**

**Data Collection:**

Data is collected from the Twitter streaming API every 5 minutes during 5 minutes on the principal actor of the French presidential election

The data is collected from Google trends about searches around the main actors.

**Preprocessing Steps:**

* Removed Retweets:

Retweets were just a burden on the system. So, with the help of Boolean value from ‘is\_retweet’ column, I removed the retweets.

* Removed Useless Columns:

I have selected only the tweet column for my processing so removed other unnecessary columns.

* Dropped Duplicates:

To ensure data quality and eliminate redundancy, I dropped duplicate rows from the dataset. This process helps in maintaining the integrity of the data.

* Removed Emojis: Emojis can often carry sentiment or meaning in text data and they are often disturbing in machine learning,So I focused on textual content. Therefore, I removed emojis from the text to simplify the data using regexp\_replace.
* Removed Mentions:

My dataset was having text tweets along with mentions, such as user names or references to other individuals,So I removed them from the text data.

* Removed URLs:

URLs present in the text data were removed to eliminate any irrelevant information and ensure consistency in our analysis. By removing URLs, we ensure that our sentiment analysis algorithms focus solely on the textual content of the tweets.

* Removed Special Characters: Special characters, such as punctuation marks and other special symbols were removed from the text data. This step helps in better analysis of the text.
* Tokenized Data:

Tokenizing the data was a important measure.So, I simply used Tokenizer function from PySpark ML function and passed my data.

* Removed Stop-words:

Stop words can be used to form a sentence but they have no contribution in sentiment analysis. So, after tokenizing the data I removed stop words from it using Stop words remover functions from PySpark ML functions.

These preprocessing steps were crucial in preparing the dataset for sentimental analysis, ensuring data cleanliness, and enhancing the effectiveness of our analysis algorithms.

**Exploratory Data Analysis (EDA):**

EDA revealed significant insights into the distribution of sentiment scores, follower demographics, and the frequency of terms in the text data. Notably, the majority of sentiments were neutral, while positive and negative sentiments were also prevalent. Furthermore, the analysis shed light on the distribution of followers among users and variations in tweet volume over time.

**Analysis Technique:**

Such techniques are:

* **NLTK:**

NLTK, or the Natural Language Toolkit, is a powerful library in Python used for natural language processing tasks, including sentiment analysis. One of the tools provided by NLTK is the SentimentIntensityAnalyzer (SIA), which calculates sentiment scores for individual tokens (words) within a text. These scores range from -1 to 1, where negative values indicate negative sentiment, positive values indicate positive sentiment, and values close to zero indicate neutrality.

* **Logistic regression:**

In sentiment analysis of textual data, logistic regression stands out as a robust technique for sentiment classification. Logistic regression, a statistical method commonly employed for binary classification tasks, is particularly well-suited for sentiment analysis. The primary objective of sentiment analysis is to categorize text into distinct sentiment categories, such as positive, negative, or neutral.

In this tweet sentiment analysis, I collected tweets using the KAGGLE and filtered them based on keywords and hashtags related to our topic of interest. We preprocessed the tweets by removing stop words, punctuation, and URLs, and emojis.We used a pre-trained sentiment analysis model based on a deep learning architecture, and displayed the final form according to the text and sentiment recognized in it.

**Findings and Interpretation:**

The sentiment analysis conducted on the dataset revealed insights into the overall sentiment expressed in the text data. By calculating sentiment scores for each text entry, we gained an understanding of the predominant sentiment trends present in the dataset.

Analysis of the sentiment scores allowed us to observe the distribution of sentiment across the dataset. This included identifying the proportion of text entries categorized as positive, negative, and neutral, providing valuable insights into the overall sentiment landscape of the dataset.

The preprocessing techniques employed, such as removing emojis, mentions, URLs, and special characters, had a notable impact on the sentiment analysis results. These techniques helped in cleaning the data, resulting in more accurate sentiment analysis outcomes.