Design and methodology

The DAG Mean Tweet Pipeline is a system that uses the content of each tweet to calculate the average sentiment score for each tweet. Several interconnected modules in the pipeline carry out a variety of tasks, including data collection, data pre-processing, sentiment analysis, and data visualization. The pipeline's design and method are outlined below.

Data Ingestion: The first stage of the pipeline involves collecting statistics from Twitter. This can be accomplished through the usage of the Twitter API, which presents access to Twitter statistics circulate. The information may be filtered based totally on keywords, hashtags, or other standards to make certain that the best relevant tweets are collected.

Data Pre-processing: Once the statistics have been accumulated, it wishes to be pre-processed to make certain that it's miles in a layout that may be analysed. This might also involve cleaning the statistics, doing away with any irrelevant facts, and transforming the information into a format that is like-minded with the sentiment analysis device. Utilizing Python libraries like NLTK, clean and pre-process the raw tweet data by removing stop words, URLs, special characters, and other noise.

Sentiment Analysis: The sentiment analysis stage entails using a gadget studying algorithm to categorise every tweet as both positive, poor, or neutral. This may be achieved through the usage of a pre-skilled model or via training a brand new version of the usage of the collected statistics.

Data Aggregation: Once the sentiment scores had been calculated for every tweet, they may be aggregated to calculate the common sentiment rating for the whole set of tweets.

Implementations-:

The execution of the DAG information pipeline for Mean Tweet examination includes the utilization of different instruments and advancements, for example,

Airflow by Apache: The data pipeline design, scheduling, and monitoring platform Apache Airflow is open-source.

Pandas: Pandas is an information control library utilized for information pre-handling and conglomeration.

Text Blob: The sentiment analysis library Text Blob is written in Python.

NTKL-: Tokenization, part-of-speech tagging, stemming, and sentiment analysis on text data can all be done through intuitive interfaces.

Matplotlib: The data visualization library Matplotlib is written in Python.

Docker: For simplicity of deployment and scalability, the entire pipeline is packaged into a single container with the help of the containerization platform known as Docker.

Retrieve Tweets: The tweet dataset is retrieved from a source, such as the API of a social media platform, in this task. The dataset can be put away in a data set or document framework for later handling. There are no dependencies on this task.

Tweets that are free of clutter: This task standardizes the text and removes irrelevant data from the raw tweet data using a variety of cleaning methods. The outcome of the Fetch Tweets task is required for this task.

Analyses: The cleaned tweets are analysed for useful information like sentiment, keywords, or mentions in this task. For this situation, we need to register the mean tweet length. The results of the Clean Tweets task are required for this task.

Find the Mean: Based on the Analyses task's output, this task calculates the mean tweet length. There are no dependencies on this task.

Text

Description automatically generated

Text

Description automatically generated

Results-:

Results of Mean Tweet Analysis Using the DAG data pipeline, it is possible to gain a better understanding of the overall sentiment trends and patterns over time. A portion of the key outcomes that could be gotten from the investigation include:

For each tweet, the average sentiment score was: This helps determine whether each tweet has a positive, negative, or neutral sentiment.

Feeling patterns after some time: This aids in distinguishing the general feeling patterns over the long run and whether some particular occasions or events are affecting the opinion.

Popular topics or hashtags: This aids in the identification of popular hashtags or topics that are influencing tweet sentiment.

Sentiment analysis of users: This makes it easier to determine how certain users feel and how they affect the sentiment of the tweets as a whole.

In general, the DAG data pipeline for Mean Tweet analysis is a powerful tool for studying sentiment trends and patterns over time in large quantities of tweets.