**Twitter Sentiment Analysis of withdrawal of US troops from Afghanistan**

**Use Case**

Twitter is one of the biggest social media platforms, with around 200 million daily active users around the world generating daily tweets in the range of half a billion, no wonder twitter has access to one of the biggest textual data. By examining these tweets, one will notice that most of the times there is a sentiment, opinion and other subjective information associated with these tweets. So, expanding the analysis of tweets to identify the sentiments from the tweets of the users from a geographic region or for a period of time, it becomes possible to assess to some extent the general opinion that people hold with respect to different issues. One common sentiment analysis is the polarity detection which classifies statements as positive, negative, or neutral. I intend to perform sentiment analysis on the tweets relating to the withdrawal of US troops from Afghanistan with the intention of figuring out the opinion that people hold for the issue.

**Architectural Diagram**

The architecture for the project will be as below

Diagram

Description automatically generated

**Data Storage Platform**

I am going to use Bigquery on google cloud platform to store my streaming data. However, before storing the data in the Bigquery, I am going to transform the semi-structured text data to a structured form using cloud function service. The data transformation will be followed by data cleansing process where the tweets will be cleaned either by discarding unwanted tweets or by improving the quality of the tweets by removing unwanted characters. The data cleaning process might also involve selection of relevant data fields coming from the source that might prove beneficial for further analysis. All the tasks relating to the data quality will be performed using composer service of the GCP.

**Data Transformation**

For my activity, the data coming in the form of JSON files in real time will be ingested into the GCP. The rate of ingestion of the data is going to be per second where every second the data is going to be fetched from twitter’s API and ingested into the GCP. However, since the data will come in the form of JSON files, I will convert them in the tabular form, perform certain cleaning activities to improve the data quality and then ingest the data into GCP. At the moment, I do not have any transformation relating to file parsing, aggregation, or any joins on data in mind to perform, however, once I am well into the project, I’ll have better control over the data and thus might perform some transformation.

**Visualization**

I do aim to build some visualization once I have access to good quality data. The visualization could be bar chart or pie chart, however, once I progress into the project, I might decide to improve the visualization. Although running an analytic algorithm on the final data is possible but I am not going to run any predictive or statistical analysis on the end data. Finally, the format of the data will be tabular to build visualization.