

CLOUD APPLICATION DEVELOPMENT

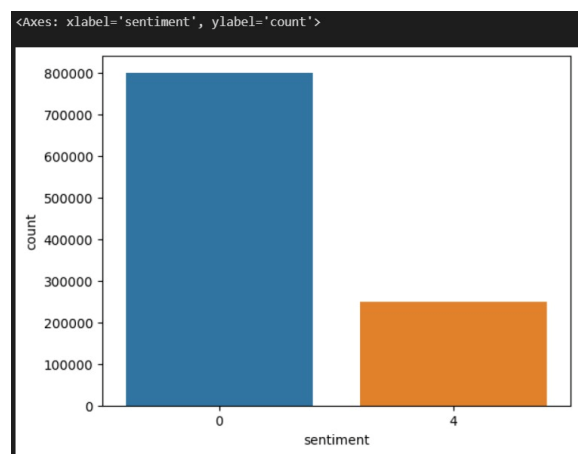
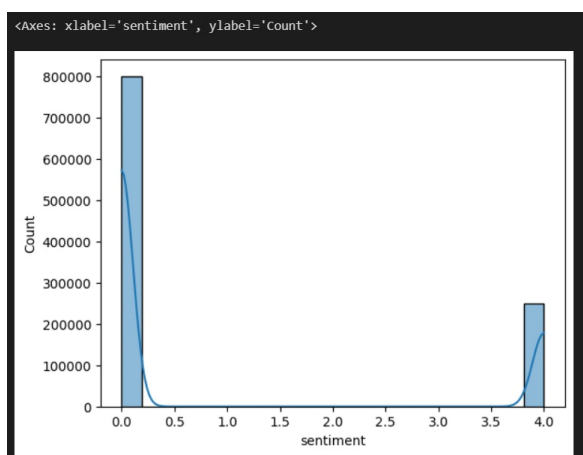
GROUP 3 – PHASE 4

INTRODUCTION:

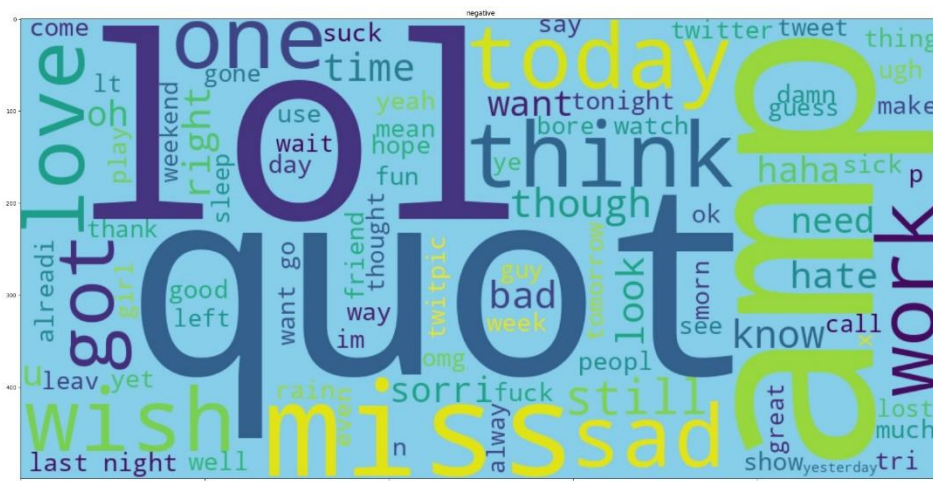
Continuing the development of a comprehensive big data analysis solution, we progress to applying advanced analysis techniques and visualizing the results. This phase delves deeper into data, employing machine learning algorithms, sentiment analysis, and time series modeling as necessary, to uncover intricate patterns and insights. Once the analysis is complete, advanced visualization tools come into play. These tools, such as data dashboards and interactive charts, help transform complex findings into clear, actionable insights. This advanced stage empowers organizations to extract valuable knowledge from their data, facilitating data-driven decision-making and a competitive edge in their respective fields.

DATA PREPROCESSING:

Data exploration and analysis are critical steps in big data analysis solutions. Large datasets are imported into IBM Cloud Databases. Depending on the specific use case, Db2 was used from IBM Cloud. Data Cleansing and preprocessing was done to handle missing values, outliers, and inconsistencies. This step ensures that the data is in a suitable format for analysis. We used python for most cases. Analyzing using EDA we can understand the data's characteristics. This involves statistical summaries, data visualization, and identifying patterns or trends. IBM Cloud offers analytics and visualization tools for this purpose. Basic data cleaning and transformation is a critical step in the data analysis process. It involves identifying and rectifying inaccuracies, inconsistencies, and missing values in the dataset. It includes tasks like removing duplicate entries, handling outliers, standardizing data formats, and inserting missing values. The objective is to ensure that the data is in a consistent, usable format, ready for analysis. Effective data cleaning and transformation enhance the quality of analysis, leading to more accurate and reliable insights from the data EDA - data analysis.



VISUALIZATION - WORD CLOUD:

[illegible]

CONFUSION MATRIX RANK AND ACCURACY:

A confusion matrix is a fundamental tool in evaluating the performance of a classification model. It's a table that compares the predicted and actual values of a classification task. The matrix comprises four values: true positives, true negatives, false positives, and false negatives, which enable you to assess the model's accuracy, precision, recall, and other key metrics. Accuracy, in particular, is a common performance metric and represents the proportion of correct predictions over the total predictions. It's a straightforward way to measure how well a model is classifying data correctly, but it may not always be the most appropriate metric, especially when dealing with imbalanced datasets. Therefore, it's important to consider other metrics alongside accuracy, like precision, recall, and F1 score, to get a more comprehensive view of a model's performance in real-world applications.

