Report

Analysis of Machine Learning Algorithm

Objective

The primary objective of this assignment was to analyze and enhance the data collected during the previous LLM task using various Machine Learning techniques. The focus was to improve chatbot response retrieval by utilizing data embeddings, clustering, probabilistic methods, linear regression methods based on user feedback and collected chunks.

Notebook and Csv files : Link

From where we collected feedback data: We extracted 163 questions from a shared Google Sheet, retrieved the best matches from the embedded data, and used the LLama (Groq API) to evaluate feedback of the relevance of the retrieved answers. Finally, we performed an analysis on the collected feedback data.

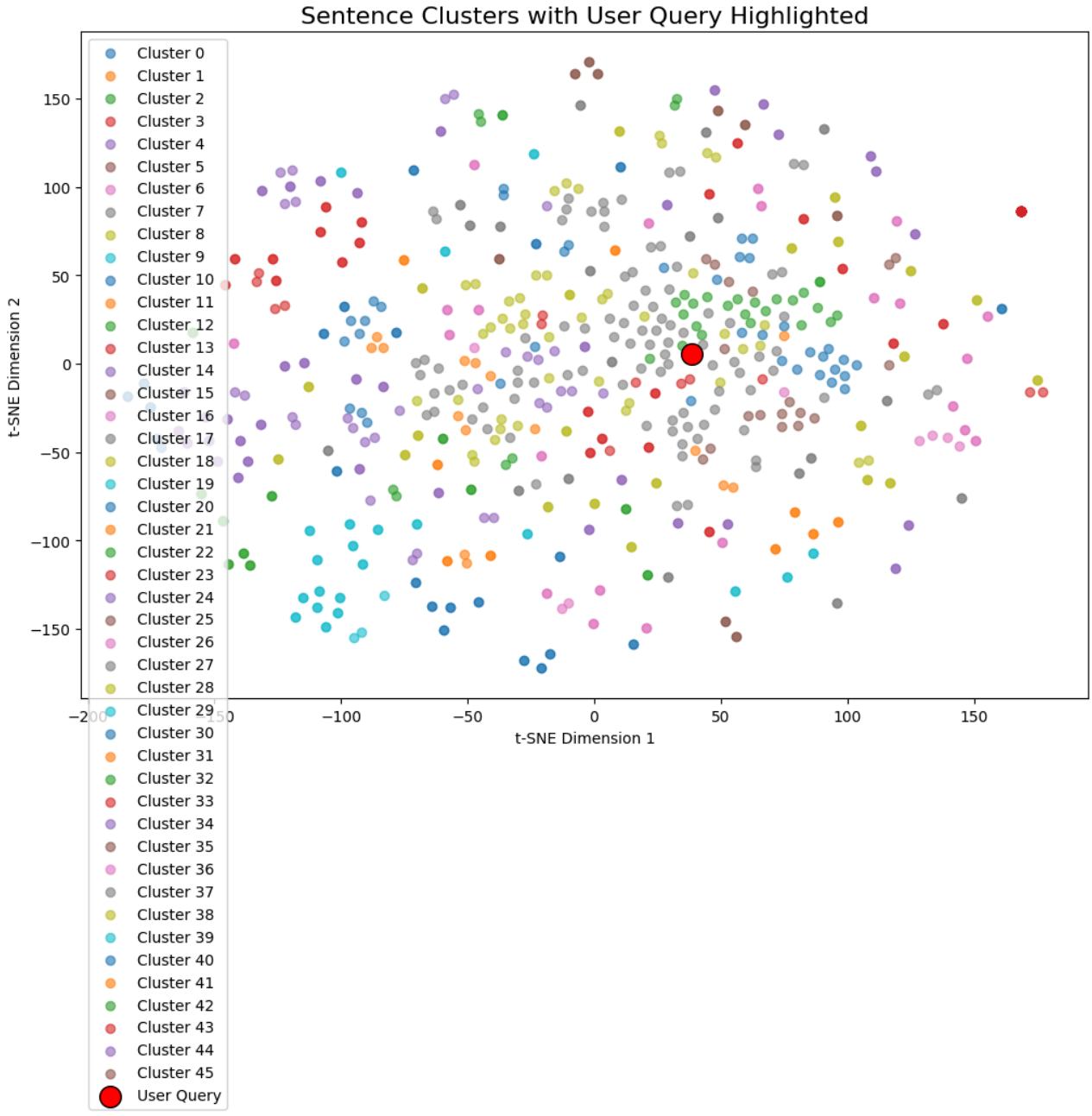
Methodology

Data Preprocessing and Embeddings

- **Data Cleaning:** The dataset was cleaned to remove noise and inconsistencies.
- **Embeddings:** Text data was converted into vector embeddings to capture **semantic** meaning, enabling more accurate query-response matching.

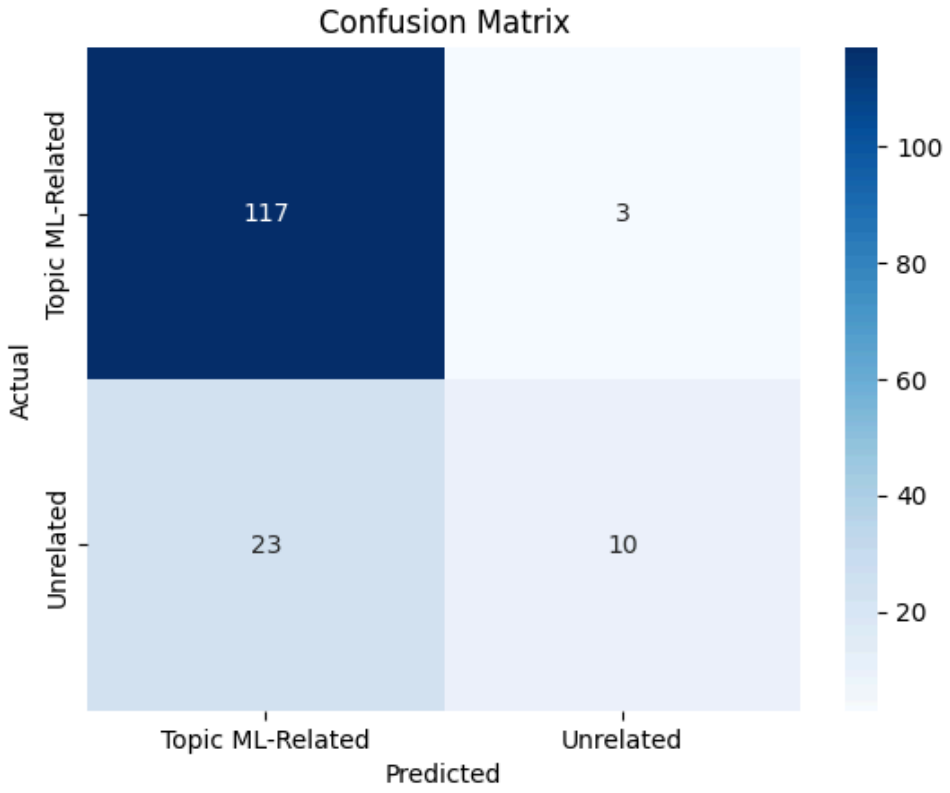
K-Means Clustering:

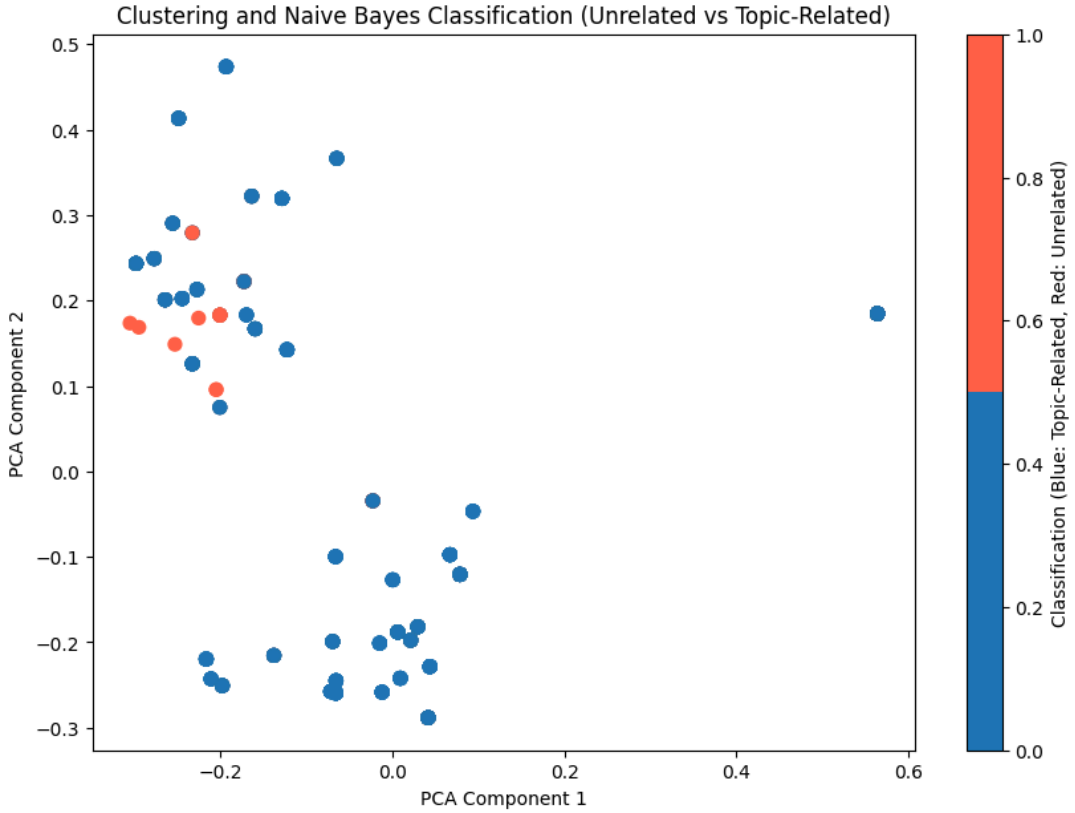
- **Purpose:** Clustering helps in grouping similar items together based on their features. In this assignment, the embeddings were grouped using the K-Means algorithm. We use embeddings of the cleaned dataset (Link).
- **How it works as per our understanding:**
    - The algorithm divides the data into a predefined number of clusters (K).
    - It assigns each data point to the cluster with the nearest centre.
    - The centres are updated iteratively to minimize the distance between data points and their respective centres.
- **Impact:** By grouping similar queries and responses, clustering improved the organization and efficiency of the retrieval process, ensuring that similar types of questions were handled consistently.



The above graph groups data into topic-based clusters with defined centers. Queries first match the closest center to identify the relevant cluster, then retrieve answers from that cluster to reduce retrieval time. Output is just below:

```
Predicted Chapter: real-life-applications-of-machine-learning
```


Confusion Matrix
Topic ML-Related
Unrelated
Actual
117
3
23
10
100
80
60
40
20
Topic ML-Related
Unrelated
Predicted

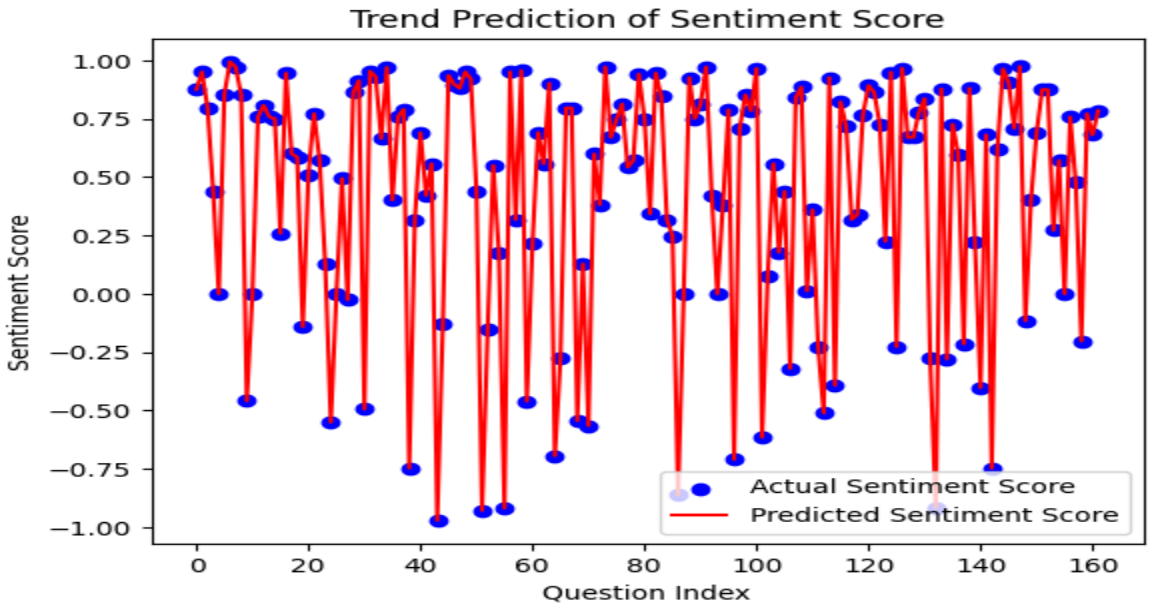

Clustering and Naive Bayes Classification (Unrelated vs Topic-Related)
PCA Component 2
PCA Component 1
Classification (Blue: Topic-Related, Red: Unrelated)

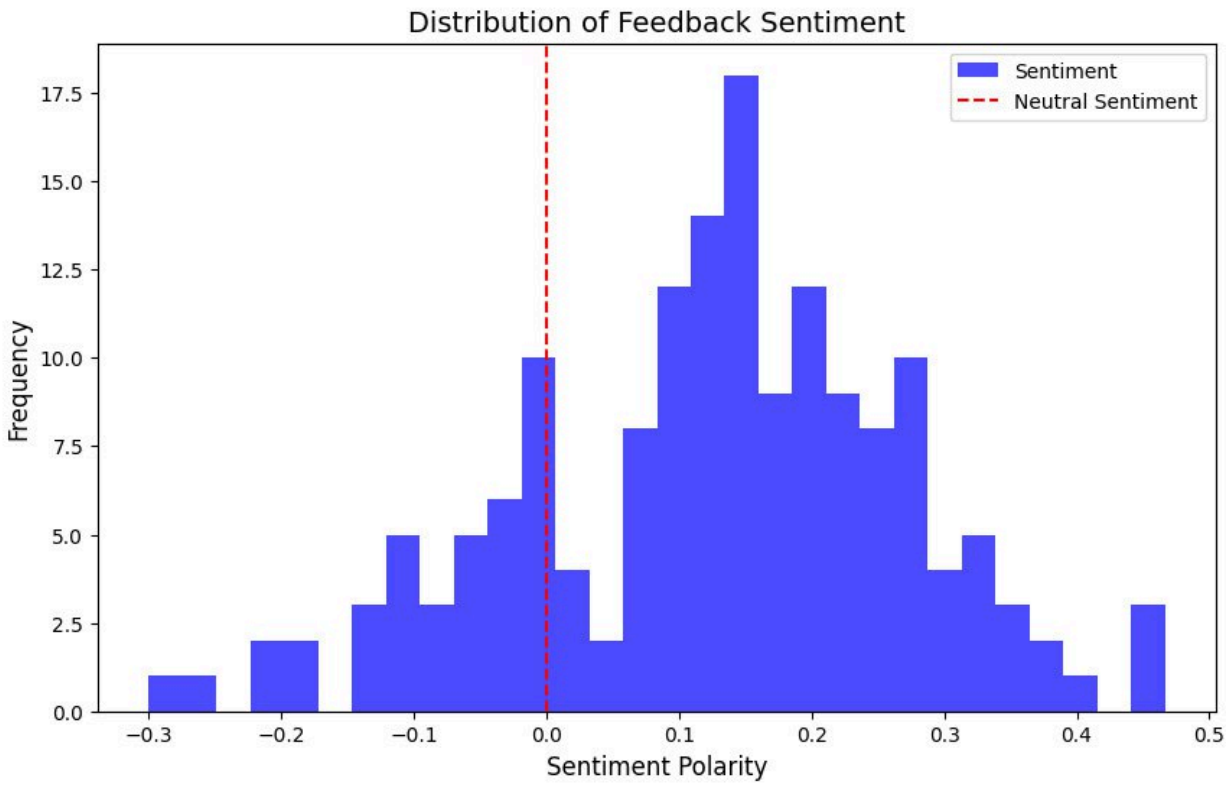
The graph shows the clustering of sentences based on reduced embeddings, with classifications indicating whether each sentence is topic-related (blue) or unrelated (red). It visualizes the Naive Bayes model's classification results on the dataset.

Sentiment Analysis (Linear Regression)

Purpose: Sentiment analysis helps in understanding user emotions and opinions from textual feedback. It categorizes feedback as positive, neutral, or negative.

- **How it works:**
  - The feedback text is preprocessed and analyzed using a sentiment analysis model.
  - The model assigns a sentiment score or label based on the words and context within the feedback.
- **Impact:**
  - Sentiment analysis provided insights into how users perceived the chatbot's responses.
  - Positive feedback indicated satisfaction, while negative or neutral feedback highlighted areas for improvement.
  - These insights were used to continuously refine and improve the chatbot's performance, ensuring that user needs were met more effectively.

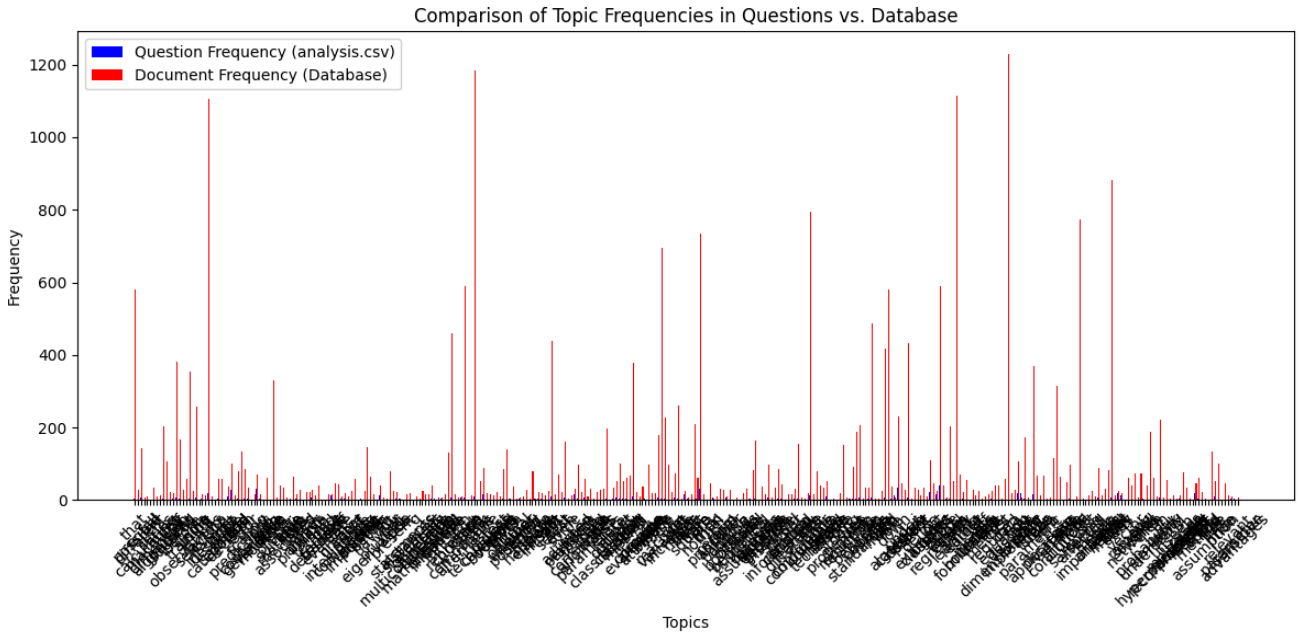




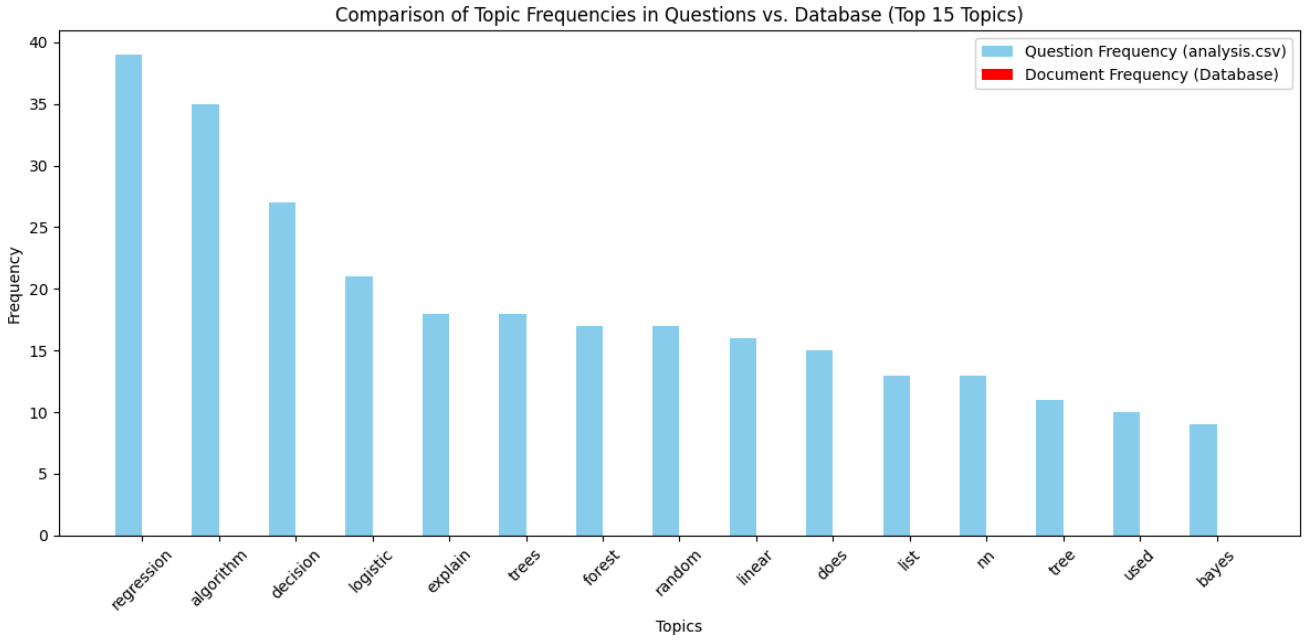
The above graphs show sentimental analysis of feedback based on asked questions and retrieve answers from the database. Closer to 1 shows completeness of answer and closer to -1 shows unsatisfactory response from database. It will help to see feedback of users in a broad view to improve data.

Visualization

- Graphs and plots of embeddings and clusters helped in understanding data relationships and improving retrieval accuracy. Some sample are shown below:



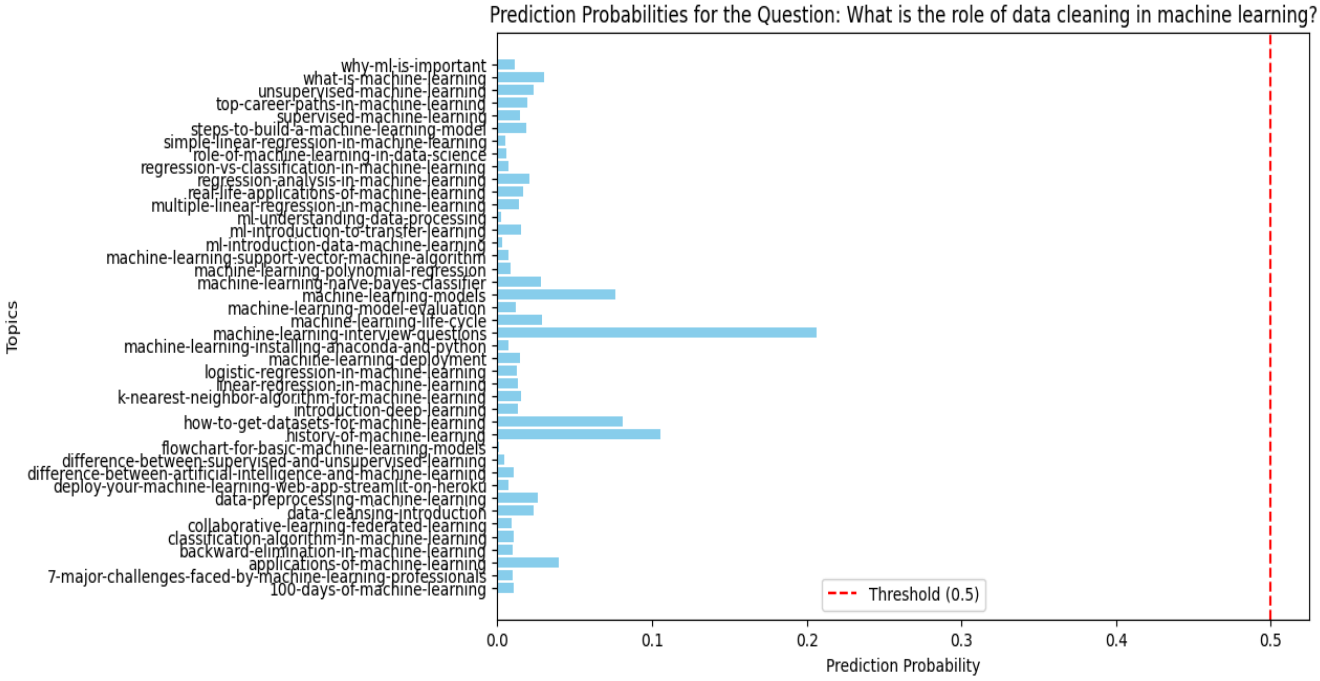
The above graph shows frequency of no of asked questions.

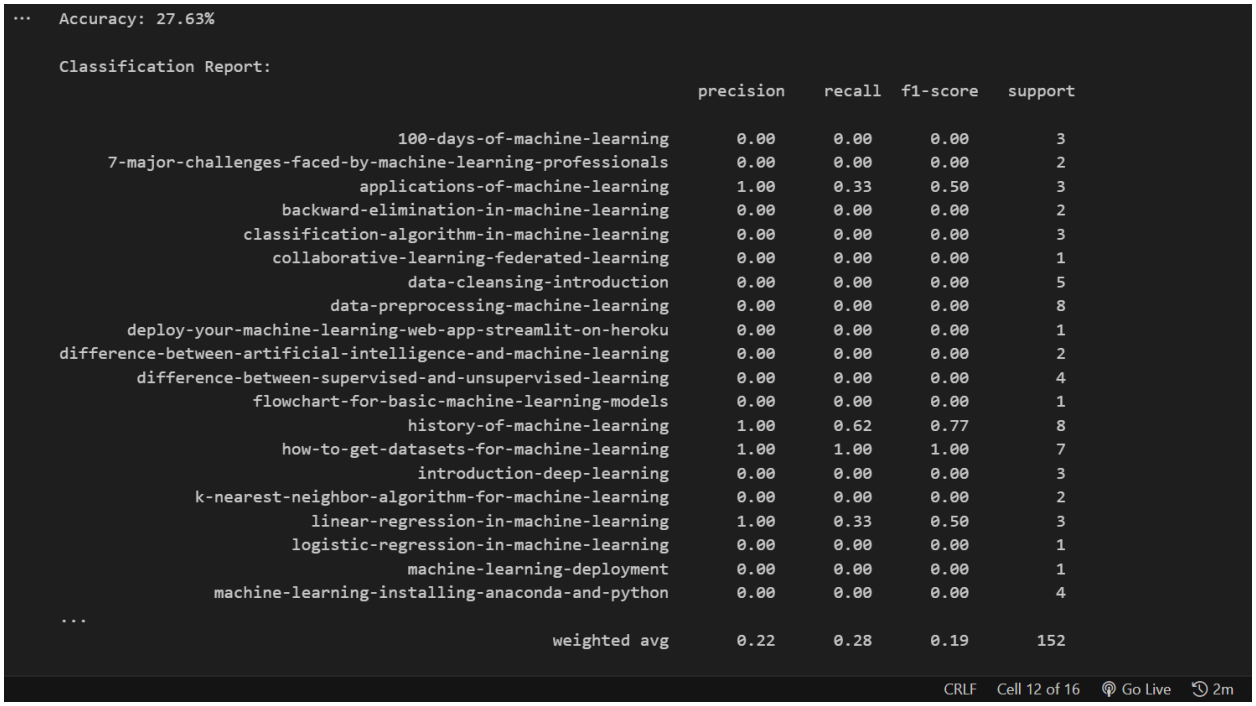


This graph shows the most frequently asked questions related to a specific topic that will help to improve our respective asked topics data.

Probabilistic Classification (Naive Bayes Theorem)

- **Purpose:** Naive Bayes is a probabilistic classifier that predicts the likelihood of a data point belonging to a specific class.
- **How it works as per our understanding:**
    - It is based on Bayes' theorem, which calculates probabilities by combining prior knowledge with new evidence.
    - The "naive" part assumes that features are independent, which simplifies calculations.
    - The classifier assigns a probability score to each possible category and chooses the one with the highest probability.
- **Impact:**
    - Feedback was categorized using Naive Bayes, helping to predict whether a given response should be classified as positive, neutral, or negative.
    - This method helped refine the system by identifying areas where user feedback suggested improvements were needed.



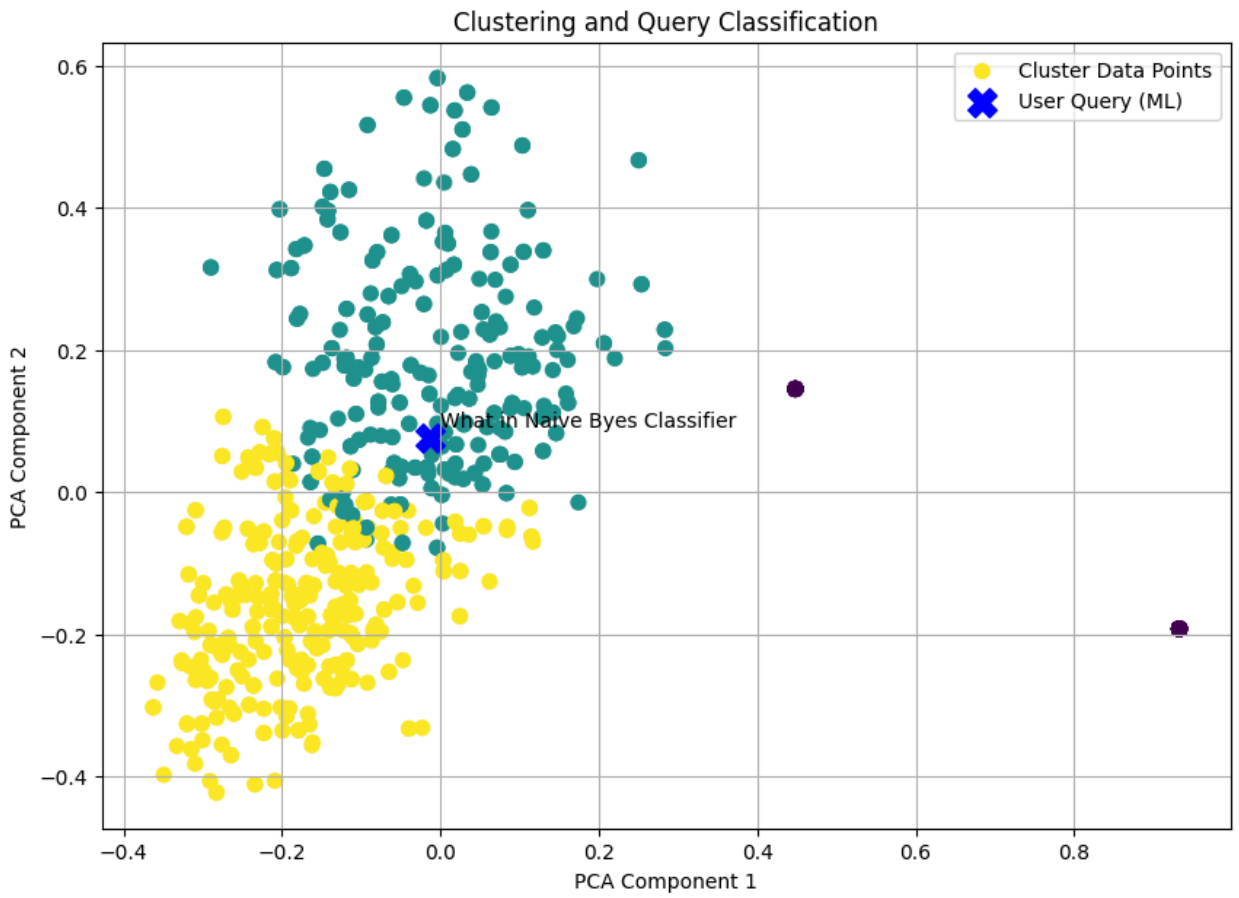


```
Accuracy: 27.63%

Classification Report:
                                                    precision    recall  f1-score   support

                                100-days-of-machine-learning       0.00      0.00      0.00         3
     7-major-challenges-faced-by-machine-learning-professionals       0.00      0.00      0.00         2
                             applications-of-machine-learning       1.00      0.33      0.50         3
                    backward-elimination-in-machine-learning       0.00      0.00      0.00         2
                  classification-algorithm-in-machine-learning       0.00      0.00      0.00         3
                     collaborative-learning-federated-learning       0.00      0.00      0.00         1
                                data-cleansing-introduction       0.00      0.00      0.00         5
                          data-preprocessing-machine-learning       0.00      0.00      0.00         8
       deploy-your-machine-learning-web-app-streamlit-on-heroku       0.00      0.00      0.00         1
difference-between-artificial-intelligence-and-machine-learning       0.00      0.00      0.00         2
        difference-between-supervised-and-unsupervised-learning       0.00      0.00      0.00         4
                  flowchart-for-basic-machine-learning-models       0.00      0.00      0.00         1
                                history-of-machine-learning       1.00      0.62      0.77         8
                   how-to-get-datasets-for-machine-learning       1.00      1.00      1.00         7
                             introduction-deep-learning       0.00      0.00      0.00         3
            k-nearest-neighbor-algorithm-for-machine-learning       0.00      0.00      0.00         2
                        linear-regression-in-machine-learning       1.00      0.33      0.50         3
                   logistic-regression-in-machine-learning       0.00      0.00      0.00         1
                                machine-learning-deployment       0.00      0.00      0.00         1
              machine-learning-installing-anaconda-and-python       0.00      0.00      0.00         4
...
                                               weighted avg       0.22      0.28      0.19       152
```


The above two graphs show the prediction probabilities of various topics (or classes) for a given question. It visualizes how confident the model is about each possible topic by plotting the probability values on a horizontal bar chart, with a red dashed line representing a threshold of 0.5 for classification. The topic with the highest probability is the predicted label.



The graph shows a collection of sentences grouped into clusters. When you enter a query, it checks if it's related to machine learning (ML). If the query is about ML, it will be shown as a blue X on the graph. If it's not related to ML, it will appear away from the clusters. The graph helps to see how the query fits with the other sentences in the clusters.

Future Work:

a) To improve retrieval efficiency, will assign cluster numbers to the main database storing embeddings. When processing a query, first match it with the nearest cluster center, then search within the corresponding cluster to find the most relevant answer.
b) Will improve the chunk quality and increase the data on most asked questions by users.
c) Will apply Naive bays analysis to fast retrieval and giving the result fast that the query asked by user is available or not in data.

Integration with Generative AI (RAG Pipeline):
The IR pipeline was successfully connected to **LLaMA 3.1 70B** generative model. This enabled RAG, allowing the system to generate concise summaries based on retrieved information. This addition enhances user experience by providing more contextually relevant answers.

Web Portal Development:
A web portal was developed, ensuring users to interact with the system by asking questions related to ML. Users can also provide feedback, like, or dislike on the generated responses. This portal helps gather data and feedback regularly, which helps improve the system and make it more accurate.

Key Learnings

- **Improved Retrieval with Embeddings:**
  Enhanced the chatbot's ability to match user queries with relevant responses based on semantic similarity.
- **Effective Categorization through Clustering:**
  Clustering responses into similar groups improved the efficiency and accuracy of information retrieval.
- **Significance of User Feedback:**
  Provided valuable insights into user experiences, emphasizing the importance of improving the chatbot's performance based on feedback.
- **Visualization for Insights:**
  Graphical representations of embeddings and clusters helped in understanding data structure and relationships for better decision-making.

Team Members & Contributions:

In this part, we all worked together rather than doing individually because it was mainly an algorithmic part.

1. **Nikhil Raj Soni & Nikhil Yadav:**
   - Worked on Clustering Algorithm to get more views
2. **Jitendra & Rashmi**
   - Worked on Linear Regression to get the trending of asked questions and more analysis.

Everyone helped in Naive Bayes Classification.