Report on the SomaBot

Introduction

The **SomaBot** is an AI-powered chatbot designed to provide intelligent responses to questions related to Kenya’s Competency-Based Curriculum (CBC). With the increasing demand for accessible and accurate information on CBC, SomaBot serves as an interactive assistant for students, parents, and teachers, offering real-time answers to frequently asked questions.

This report provides an in-depth overview of SomaBot’s development, functionalities, and implementation. It explores the chatbot's architecture, natural language processing (NLP) models, and training data sources, including CBC FAQs and social media discussions on twitter. Additionally, the report discusses fine-tuning strategies used to enhance response accuracy, deployment considerations, and potential improvements for future iterations.

By leveraging advanced AI techniques, including deep learning and transformer-based models, SomaBot aims to bridge the information gap in Kenya’s education system, making CBC-related knowledge more accessible and user-friendly.

Project Overview

This project involves building a chatbot to provide accurate and real-time responses to questions related to Kenya’s Competency-Based Curriculum (CBC). It aims to assist students, parents, and teachers by offering structured and also reliable information including curriculum details, assessments, and policies. It enhances engagement, provides instant information, and simplifies learning pathways using AI-driven conversational support tailored to user needs.

Background

CBC in Kenya has introduced a new approach to education, emphasizing skill development, creativity, and critical thinking. However, many students, parents, and teachers face challenges in understanding and adapting to the new system. With frequent policy updates, complex assessment methods, and varied interpretations of CBC guidelines, there is a growing need for an accessible and reliable source of information.

Business Understanding

The Competency-Based Curriculum (CBC) in Kenya is designed to equip learners with practical skills, but its implementation has faced challenges. Parents, teachers, and students often struggle to access accurate and timely information regarding the curriculum, assessment methods, and their respective roles. Currently, most CBC-related inquiries rely on government circulars, school meetings, or online discussions, which are often fragmented and inconsistent.

To address this, an AI-powered chatbot will be developed to provide instant, reliable, and structured responses to CBC-related questions. This chatbot will serve as an interactive platform where users can seek information about curriculum structure, assessment criteria, parental involvement, and available teaching resources. The goal is to enhance accessibility, reduce misinformation, and improve user engagement by leveraging natural language processing (NLP) for dynamic, intelligent responses.

Objectives

1. To develop a chatbot, that is SomaBot, that leverages TF-IDF, Cosine Similarity and RASA model to match user queries with relevant responses and optimize the best possible answers, ensuring accurate and efficient query resolution .

2. To analyze the question column in the FAQS dataset to identify the most frequently asked queries and common keywords, enabling optimization of chatbot responses and improving user query resolution.

3. To apply an NLP-based sentiment analysis approach using VADER and TextBlob to classify CBC-related tweets as positive, negative, or neutral, providing strategic insights into the education system.

4. To analyze sentiment trends in tweets about the CBC education system, providing actionable insights to support strategic decision-making for policy development, curriculum improvements, and stakeholder engagement

5. To build and evaluate predictive models (Random Forest, XGBoost, and Logistic Regression) for sentiment classification, comparing their performance in accurately classifying CBC-related tweets.

Scope

SomaBot is a bilingual AI chatbot designed to provide accurate and contextual responses to queries related to Kenya's Competency-Based Curriculum (CBC). The chatbot supports both English and Swahili, Query Handling and is designed for students, parents, and educators to access CBC-related information efficiently.

Business Overview

SomaBot is an AI-powered bilingual chatbot designed to assist students, parents, and educators by providing accurate and timely responses to queries about Kenya's Competency-Based Curriculum (CBC). Built using Rasa and leveraging TF-IDF with cosine similarity, it enhances educational accessibility by offering real-time information in both English and Swahili. SomaBot aims to reduce the burden on teachers, improve student engagement, and streamline curriculum understanding. The chatbot is positioned as a scalable solution for integrating AI into education, addressing gaps in information delivery while ensuring reliability and inclusivity in learning support.

Problem Statement

Accessing accurate and timely information about Kenya's Competency-Based Curriculum (CBC) remains a challenge for students, parents, and educators. Traditional sources, such as government websites and educational institutions, are often complex, difficult to navigate, or lack real-time engagement. Additionally, language barriers between English and Swahili further hinder accessibility. SomaBot addresses these challenges by providing an AI-powered, bilingual chatbot capable of answering CBC-related queries with precision and also ensures that users receive relevant responses.

The business problem

Access to accurate CBC information in Kenya is limited, leading to confusion among parents, students, and teachers. SomaBot addresses this by providing a bilingual AI-powered chatbot for real-time, reliable, and interactive CBC-related support.

Metrics of Success

1. The ideal sentiment classification model should attain a minimum F1 score of 60%.
2. The top-performing chatbot model should achieve at least 90%.
3. To successfully classify tweets into different sentiments (Neutral, Positive, Negative).

Data Understanding

To ensure the chatbot delivers relevant and accurate responses, the data was gathered from multiple sources. Official curriculum documents from the Kenya Institute of Curriculum Development (KICD) serve as the primary data source, supplemented by Ministry of Education policies, circulars, and guidelines. The data is primarily text-based, comprising of structured content from official documents and unstructured conversations from forums and user queries.

There are two datasets :

- FAQs (Frequently Asked Questions) dataset
- Tweets dataset(scrapped tweets from X, formally known as twitter)

The FAQs dataset contains 2533 entries(rows) and 2 columns, the Question and Answer columns . The tweets dataset contains 1086 rows and 6 columns, the tweet_count, username, text, created at, retweets and likes.

The tweet dataset contained CBC tweets from the January 2017 to February 2025.

Summary of features in the datasets :

✔ **FAQs dataset**

- **Question :** The user’s inquiry or the asked question.
- **Answer :** The predefined response corresponding to the question.

✔ **Tweets dataset**

- **Tweet_count:** The total number of tweets collected in a dataset or tracking count of tweets in a conversation/thread.
- **Username:** The Twitter handle of the person who posted the tweet.
- **Text:** The actual content of the tweet, including any hashtags, mentions or links
- **Created at:** The timestamp when the tweet was posted,that is, year, date, day, time
- **Retweets:** The number of times this tweet has been retweeted (shared by others).
- **Likes:** The number of likes (previously "favorites") the tweet has received.

Methodology

Data collection

This step involves gathering raw data from various sources relevant to the chatbot.

✔ **Sources of Data:**

- FAQs Dataset – Structured question-answer pairs.
- Tweets Dataset – Extracted tweets for analysis from X, formally known as twitter.

- Official Curriculum Documents – Kenya Institute of Curriculum Development (KICD) policies, Teachers Arena Document, National Curriculum Policy, KNEC, Teacher ac, Darasa app, Citizen Digital, KDRIDP
- User Queries – Data from past interactions to improve chatbot responses.

✔ Collection Methods:

- Web Scraping – Extracting FAQs from websites and tweets from X(formally known as twitter)
- Manual Curation – Collecting FAQs from government documents.

Data preparation

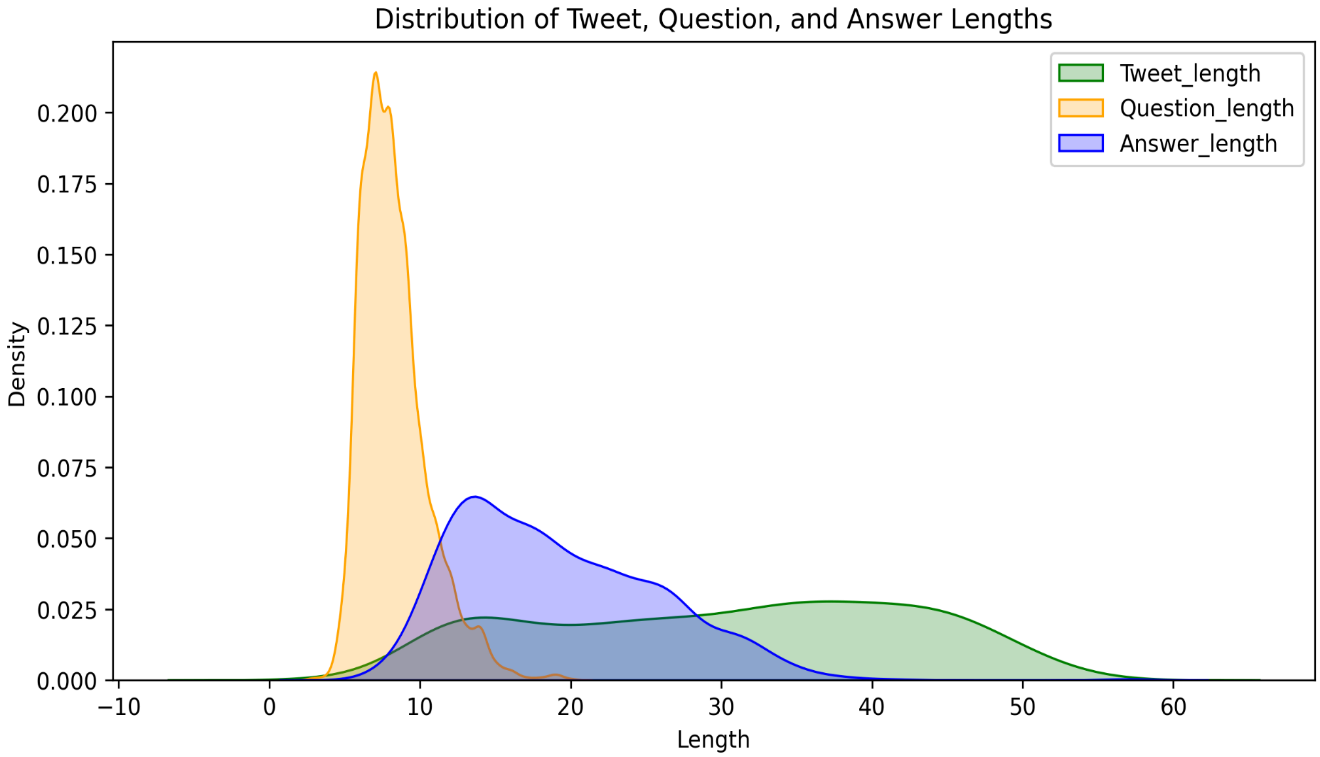
Import all the necessary libraries.
Load the datasets using pandas library.
Check and clean for missing values, outliers,duplicates in both datasets.
Dropped irrelevant columns.
Tokenization – Splitting sentences into words for processing.
Lemmatization – Converting words to their root forms
Vectorization – Converting text into numerical format.

Modelling approach

1.Exploratory Data Analysis(EDA) : Identified trend and relationships.
2.Machine Learning models : Tested Random Forest, Logistic Regression, XG Boost,TF-IDF and Cosine similarity.
3.Hyperparameter tuning : Used GridSearch CSV for optimization.
4.Rasa : An open-source conversational AI framework for building chatbots and virtual assistants.

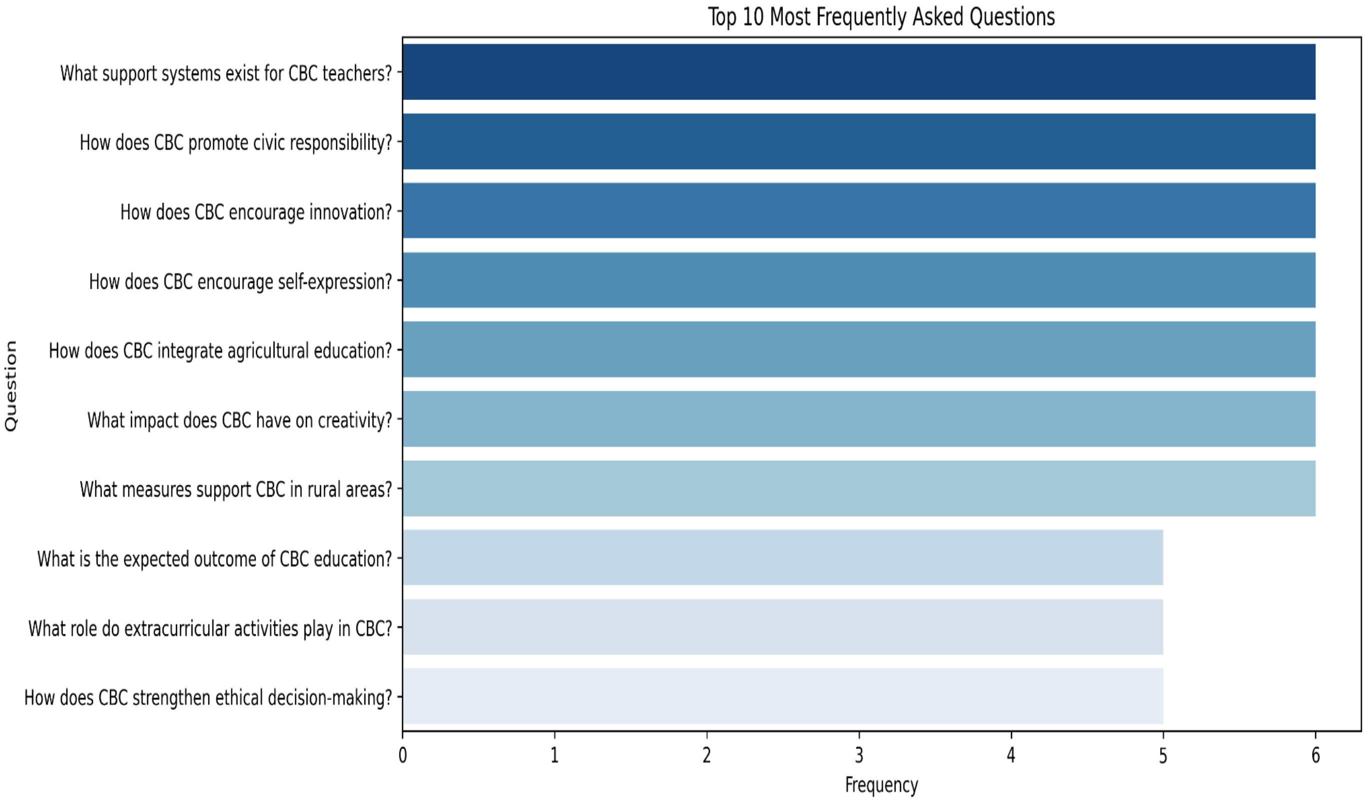
1.Exploratory Data Analysis(EDA)

Checking the distribution of the tweets length, question length and answer length



- Question Length (Orange): Most questions are short, with a peak below 10 words and few long questions. This may be so because they need to be precise and to the point.
- Answer Length (Blue): Answers vary in length, peaking at 10–20 words, with some exceeding 30 words. The Answers are more variable, suggesting that responses depend on the complexity of the question.
- Tweet Length (Green): Tweets have a broad distribution, indicating significant variation, with some being more detailed, which is expected since tweets can be concise or detailed.

To analyze the question column in the FAQS dataset to identify the most frequently asked queries and common keywords



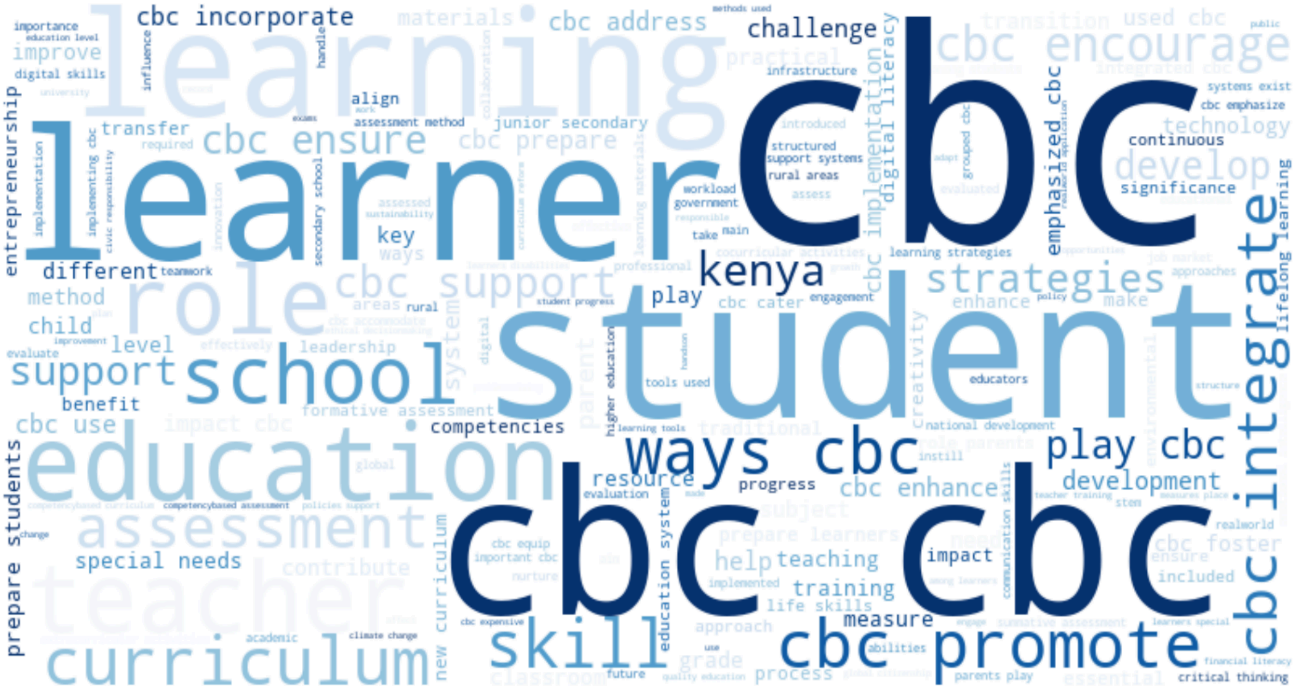
The graph displays the top 10 most frequently asked questions about CBC (Competency-Based Curriculum). Some of the most common queries are :

- What support systems exist for CBC teachers?
- How does CBC promote civic responsibility?
- How does CBC encourage innovation?
- How does CBC encourage self-expression?

The most common inquiries focus on self-expression, agricultural education, rural support, creativity, civic responsibility, innovation, teacher support, extracurricular activities, ethical decision-making, and financial literacy, highlighting key concerns in CBC implementation.

Displaying the common words from the questions column

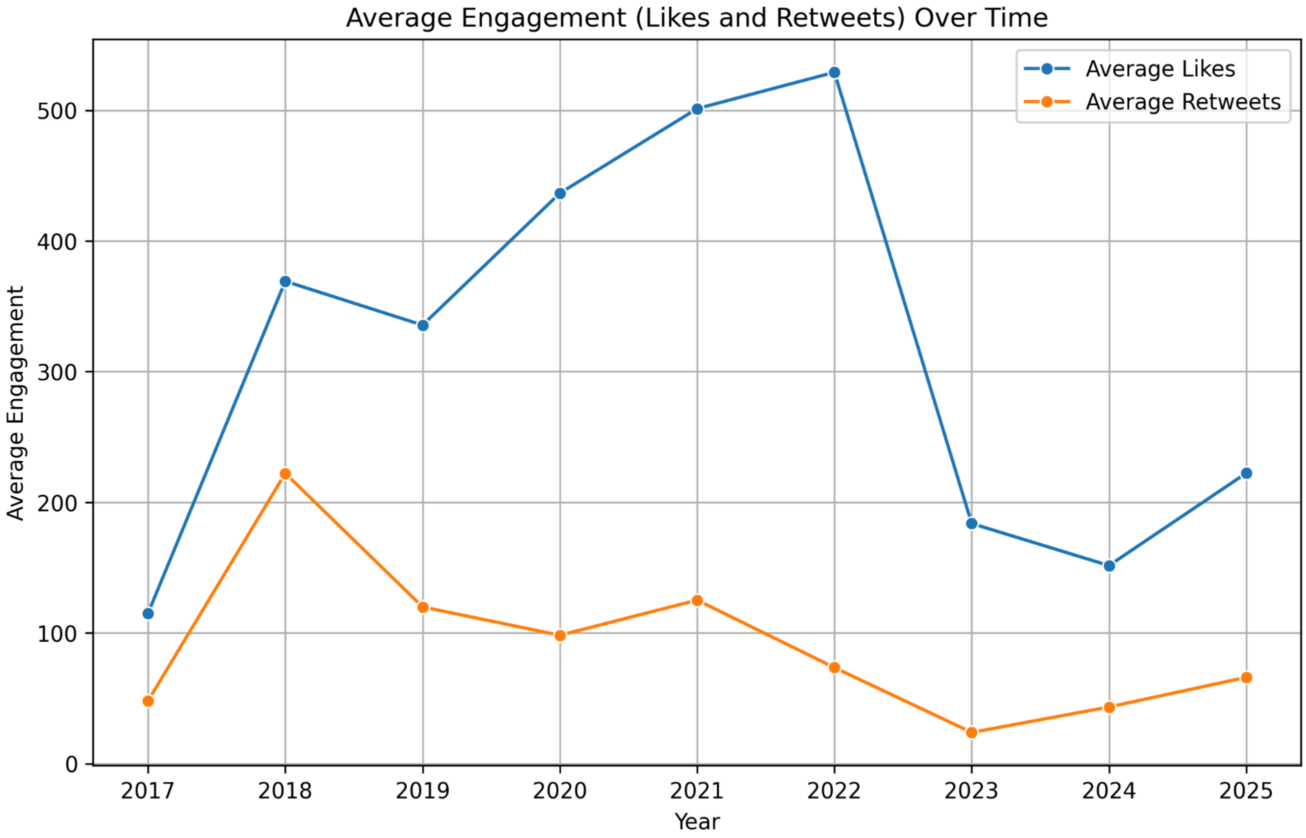
WordCloud of the FAQs Dataset



The word cloud visualization highlights the most frequently occurring words in the FAQs dataset related to CBC (Competency-Based Curriculum). The larger and bolder the word, the more often it appears in the dataset.

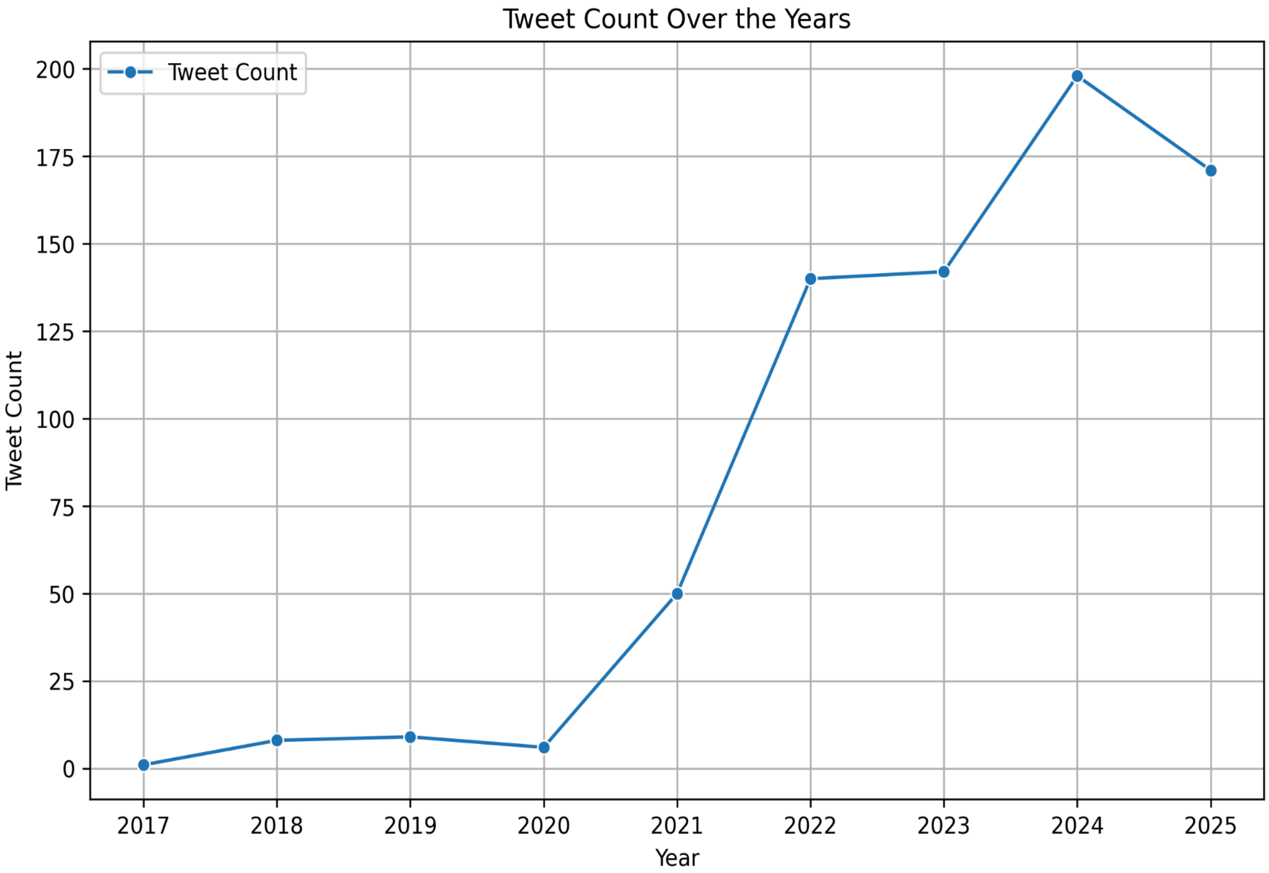
- The most prominent words are "CBC," "learning," "learner," "student," "education," and "teacher"—suggest that discussions mainly focus on students, educators, and learning strategies.
- Other words like "curriculum," "promote," "role," "skill," "grade," "level," and "support" suggest that the FAQs focus on curriculum development, skill enhancement, and educational support within CBC.

To Identify tweets engagement patterns over the years



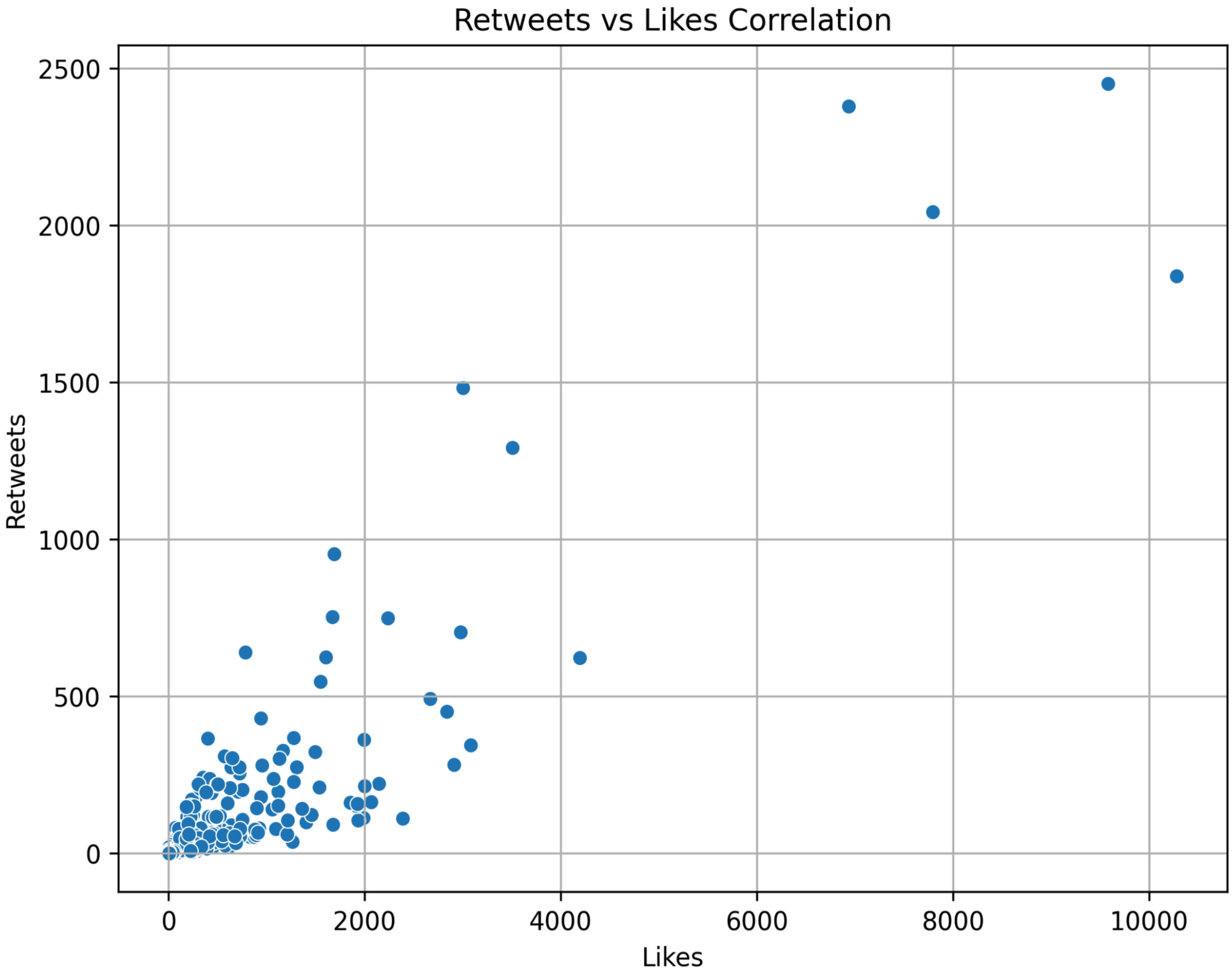
The graph shows average engagement (likes and retweets) over time from 2017 to 2025. Likes peaked around 2022, followed by a decline, while retweets peaked in 2018 and decreased significantly. Engagement dropped post-2022 but showed slight recovery in 2025.

To display the number of tweets over the years



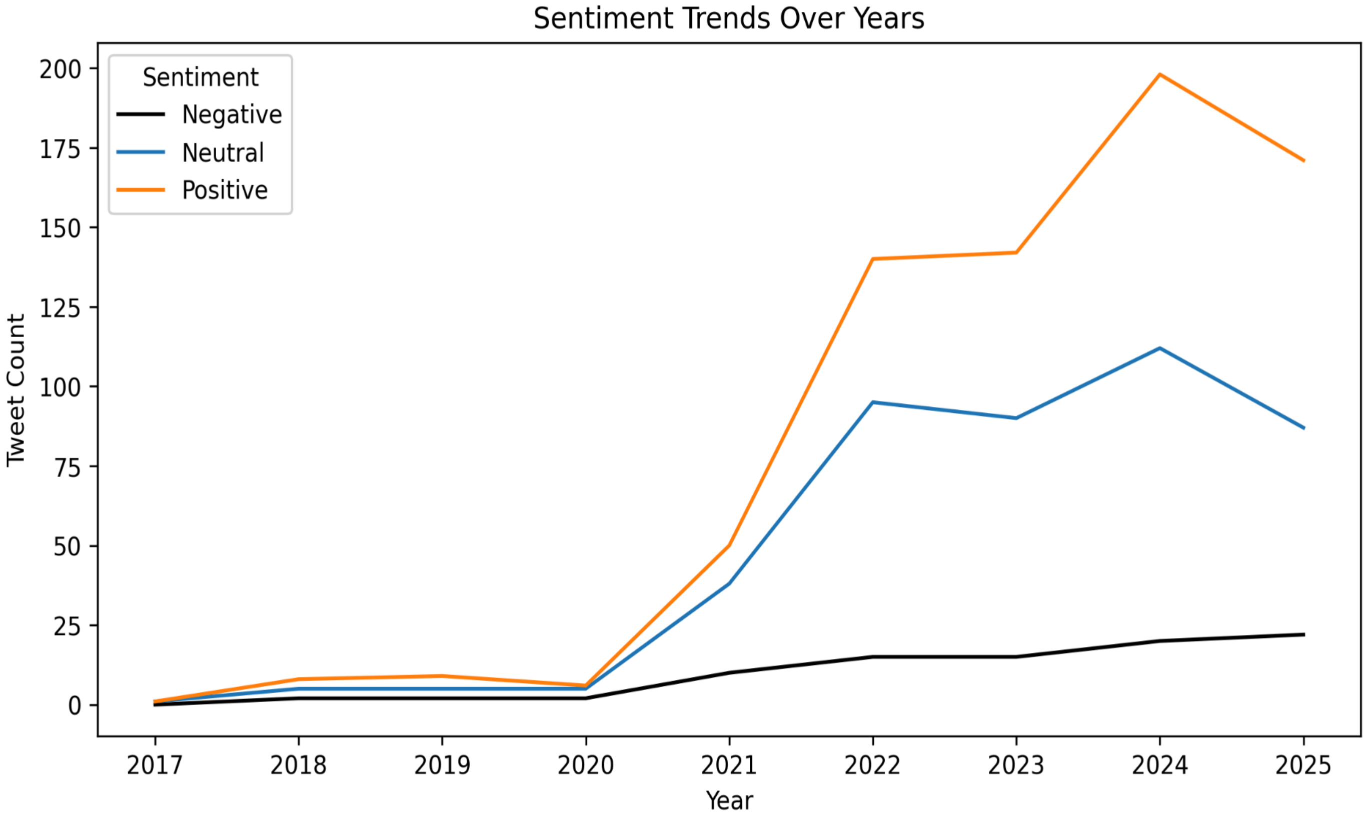
The graph represents tweet count over the years (2017-2025). Tweet activity remained low until 2020, then surged significantly, peaking in 2024. A slight decline is observed in 2025, but overall, engagement remains high.

To examine the relationship between retweets and likes on tweets



Correlation between Likes and Retweets: 0.9175776970837106
There is a high correlation of 0.92 between likes and retweets highlighting that there is a strong relationship between retweets and likes. This suggests that tweets with more retweets tend to also receive more likes, implying that content with high engagement in one metric often performs well in the other.

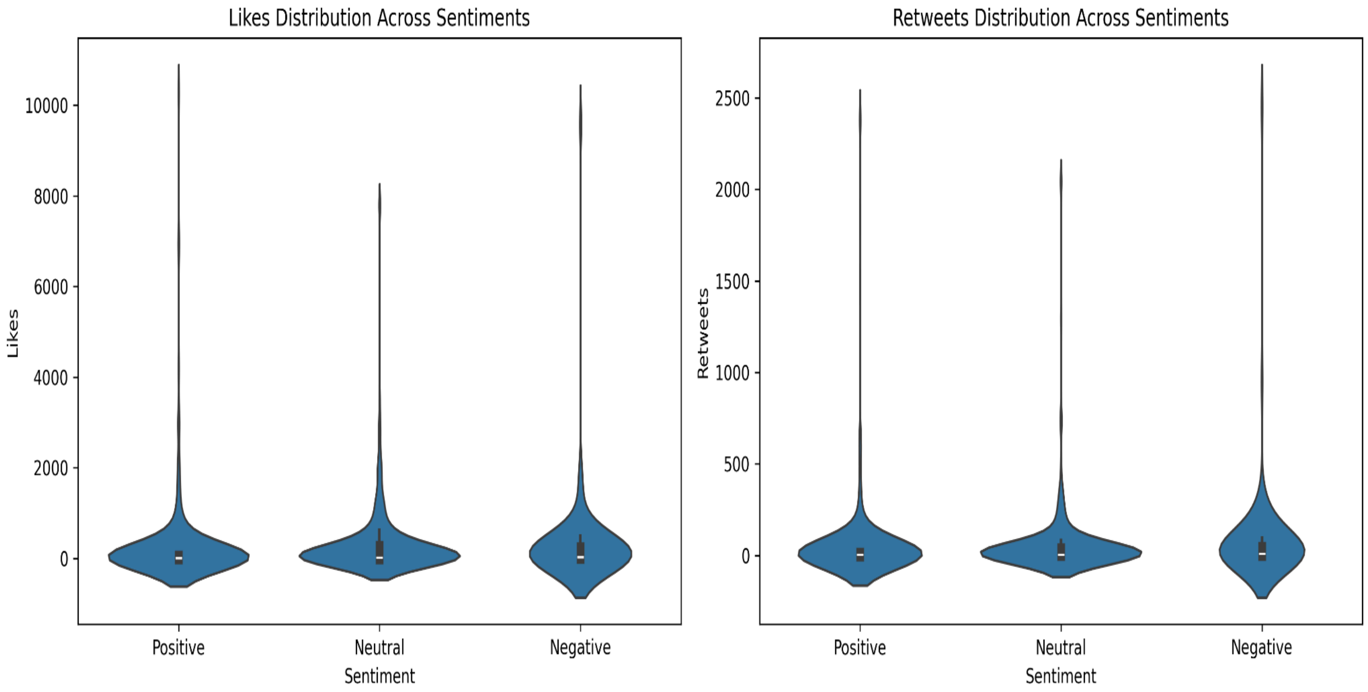
To analyze sentiment trends in tweets about the CBC education system



The sentiment trend analysis of CBC-related tweets from 2017 to 2025 shows a significant rise in positive and neutral sentiments starting around 2021, with positive sentiment peaking in 2024 before slightly declining. This suggests growing public engagement, possibly due to policy reforms or increased awareness.

Neutral sentiment follows a similar upward trend, indicating widespread discussions without strong opinions. Negative sentiment remains low but gradually increases, suggesting some concerns, though not as prominent as positive reactions. Overall, the data indicates increasing public discourse around CBC, with a generally favorable perception over time.

To determine which sentiments gets more engagement



Likes Distribution

- The majority of tweets, regardless of sentiment, receive fewer likes (clustered around 0–500 likes).
- A few tweets have extremely high likes (outliers above 10,000), which suggests that some CBC-related tweets gained viral attention.
- The distribution is similar across all sentiments, meaning no clear preference for engagement based on sentiment alone.

Retweets Distribution

- Most tweets receive low retweets (clustered around 0–100).
- Some tweets have high retweet counts (above 2,500), indicating a few highly shared tweets.
- Again, the distribution across sentiments is quite similar, suggesting no strong sentiment bias in retweet activity.

Overall, sentiment does not seem to significantly affect engagement, as distributions look quite similar across Positive, Neutral, and Negative categories.

Data preprocessing

In this section, we preprocess the data to prepare it for modelling.In the dataset we have sentiments; Neutral(1) - 350, Positive (2)- 287 and Negative(0) - 88. To tackle this, we will use SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples for class 0 (Negative) so all classes are balanced.

Step 1: SMOTE - used to handle class imbalance problems by oversampling the minority class with replacement
Step 2: Conduct a train test split on the tweets data
Balanced class distribution: Counter({1: 281, 0: 281, 2: 281})

Creating our models

Random Forest:

```
           precision    recall   f1-score   support

       0       1.00       0.05      0.09        22
       1       0.58       0.77      0.66        69
       2       0.58       0.57      0.58        54

accuracy                            0.59       145
macro avg      0.72      0.46       0.44        145
weighted avg   0.65     0.59       0.54        145
```

XGBoost:

```
          precision    recall    f1-score   support

      0      0.38        0.14      0.20        22
      1      0.63        0.71      0.67        69
      2      0.59        0.65      0.62        54

accuracy                            0.60       145
macro avg      0.53       0.50      0.50       145
weighted avg  0.58      0.60      0.58       145
```

Logistic Regression:

```
              precision  recall  f1-score  support

· · ·
accuracy                         0.59        145
macro avg       0.53       0.48      0.47         145
weighted avg  0.56       0.59      0.56         145
```

Random Forest
This model performs well for class 1 (Neutral) with a 77% recall, meaning it correctly identifies most neutral tweets but poorly predicts the Class 0 (Negative), with only 5% recall, meaning most negative tweets are misclassified.
Overall, the accuracy is 59%, and the macro-average F1-score is low (0.44) due to the poor performance on class 0.

XGBoost
This model Performs better than Random Forest, achieving 63% accuracy and better balance across all classes.However, Class 0 (Negative) still has low recall (23%), meaning many negative tweets are misclassified.
Overall: The macro-average F1-score (0.55) shows a more balanced performance compared to Random Forest.

Logistic Regression
This model Performs similarly to Random Forest, but slightly better for class 1 (Neutral) and class 2 (Positive). It however displays a Very poor recall for class 0 (Negative) at 9%, meaning it struggles to identify negative tweets.
Overall: With 58% accuracy, it is slightly worse than XGBoost, and the macro-average F1-score (0.46) shows imbalance.

In all the 3 models XGBoost is the best so far with the highest accuracy & a balanced recall.

Hyper parameter tuning

We use GridsearchCV to find the best parameters for each model.

```
Best Random Forest Model:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        22
           1       0.61      0.77      0.68        69
           2       0.59      0.63      0.61        54

    accuracy                           0.60       145
   macro avg       0.40      0.47      0.43       145
weighted avg       0.51      0.60      0.55       145
```

```
Best XGBoost Model:
              precision    recall  f1-score   support

           0       0.42      0.23      0.29        22
           1       0.67      0.80      0.73        69
           2       0.71      0.67      0.69        54

    accuracy                           0.66       145
   macro avg       0.60      0.56      0.57       145
weighted avg       0.65      0.66      0.65       145
```

```
Best Logistic Regression Model:
              precision    recall  f1-score   support

           0       0.40      0.09      0.15        22
           1       0.58      0.72      0.65        69
           2       0.61      0.61      0.61        54

    accuracy                           0.59       145
   macro avg       0.53      0.48      0.47       145
weighted avg       0.56      0.59      0.56       145
```

Random Forest F1-score: 0.4289
XGBoost F1-score: 0.5694
Logistic Regression F1-score: 0.4681

XGBoost

XGBoost performs the best among the models, achieving an F1-score of 0.5352. It balances precision and recall well, particularly for class 1 (Neutral) and class 2 (Positive).

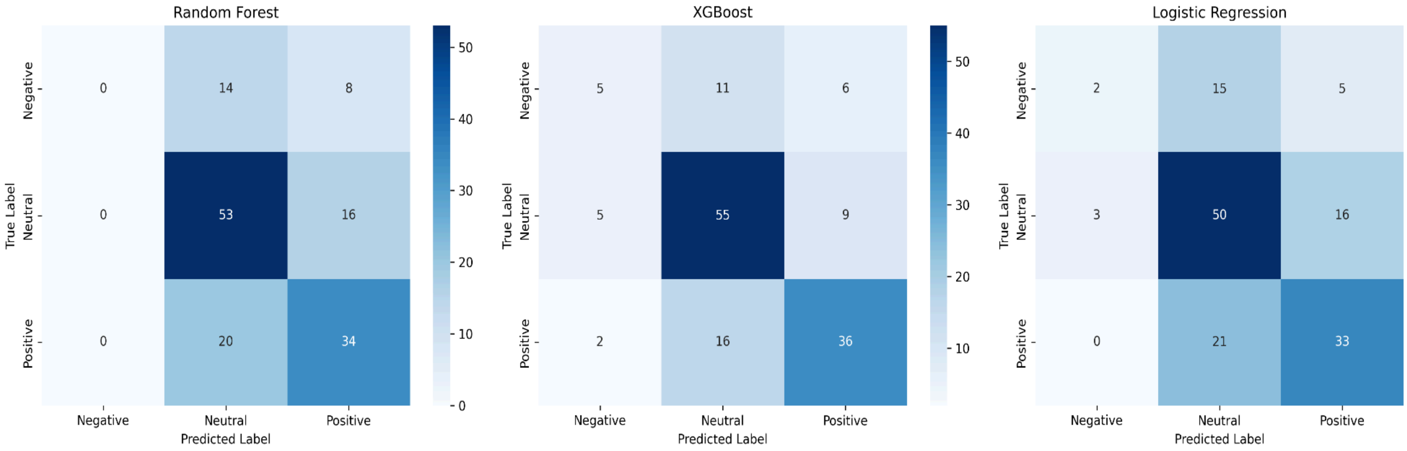
Random Forest

The model struggles with classifying class 0 (Negative) correctly, leading to an F1-score of 0.4289.
Its recall is highest for class 1 (Neutral), meaning it predicts neutral tweets more accurately than other classes.

Logistic Regression

The model performs moderately, with an F1-score of 0.4633, struggling with class 0 (Negative).
It achieves relatively stable performance across classes but lacks the predictive strength of XGBoost

Models Evaluation



Interpretation

1. Random Forest: This model struggles to classify Negative cases correctly, predicting all as Neutral or Positive. It performs well on Neutral and Positive cases but misclassifies some Positive cases as Neutral.

2. XGBoost: This model shows better balance, with more accurate classifications for Neutral and Positive cases. However, it still misclassifies some Negative cases as Neutral or Positive.

3. Logistic Regression: This model struggles the most, misclassifying many Negative cases and showing lower accuracy for Neutral and Positive predictions compared to XGBoost.

XGBoost presents to be the best-performing model for predicting CBC tweets, achieving a better balance in classifying all sentiment categories. Random Forest performs well for Neutral and Positive sentiments but struggles with Negative cases, while Logistic Regression has the highest misclassification rate across all categories.
In terms of the F1-score, XGBoost remains the top performer with an F1-score of 0.5352, indicating a stronger balance between precision and recall compared to the other models.

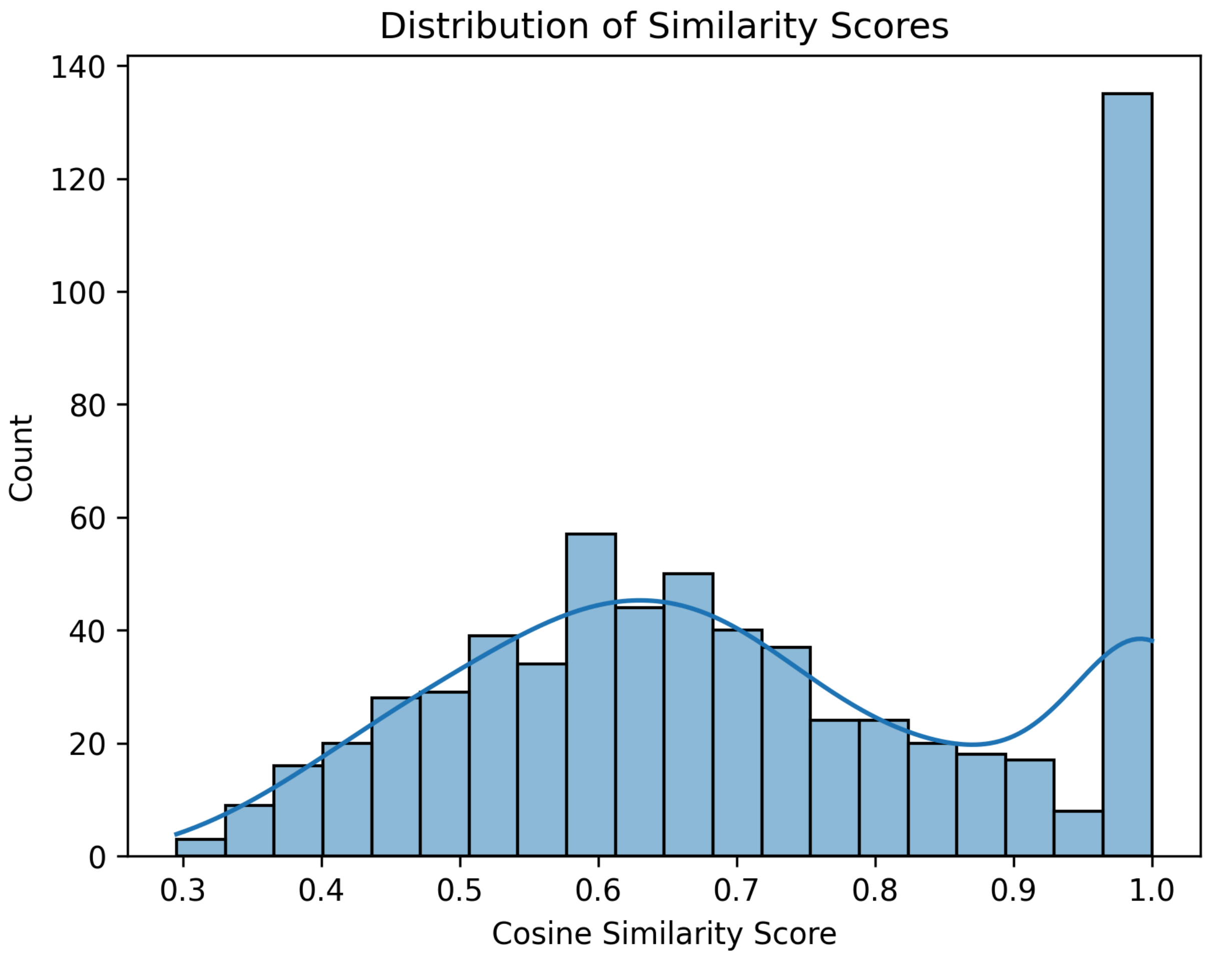
Building the model using TF-IDF + Cosine Similarity Approach

TF-IDF (Term Frequency-Inverse Document Frequency) converts text into numeric vectors, while Cosine Similarity measures the similarity between query and response, helping match user queries with relevant responses in a chatbot.

Accuracy: 0.01
Precision: 0.01
Recall: 0.01
F1 Score: 0.01
Mean Reciprocal Rank (MRR): 0.01

An MRR of 0.01 indicates very poor performance, meaning the correct answers usually appear far down the ranking list. Ideally, an MRR close to 1.0 suggests the correct answer is often the top result. To improve this score, we would enhance preprocessing and adjust the similarity threshold.

Cosine Similarity Distribution



The distribution suggests that while the model identifies some relevant matches, the sharp peak near 1.0 indicates potential overfitting. The weak matches in the 0.3–0.5 range highlight the need for a higher confidence threshold. To improve accuracy, adjusting the threshold to around 0.6–0.7 for partial matches would lead to more reliable predictions.

RASA

In the process of developing the Rasa chatbot, the location of the data folder was modified to improve the organization of resources and ensure the model's scalability. After this modification, it was essential to update the necessary paths and directories to maintain smooth functionality of the Rasa framework.

The steps followed to ensure the successful integration of the new data folder are as follows:

1. **Updating the Data Folder Path:**
   The first step was to move the data folder to a new location. This included files such as nlu.yml, stories.yml, domain.yml, and rules.yml, which form the foundation of the chatbot's training.
2. **Adjusting Configuration Files:**
   Once the folder was moved, it was necessary to ensure the Rasa configuration reflected the change. In the config.yml and other relevant files, the paths to the training data were updated. The data parameter in the terminal was also adjusted to point to the new location.
3. **Training the Model with New Data Folder:**
   With the data folder relocated and configuration files updated, the chatbot model was retrained. The following command was executed to train the Model :

   ```
   rasa train
   ```

   This ensured that Rasa accessed and processed the correct datasets during training.
4. **Verifying the New Setup**:
   After training the model, tests were conducted to verify that the chatbot was functioning as expected. This involved running the chatbot in the shell and checking if all the intents, actions, and responses were correctly processed. The following command was run:

   ```
   rasa shell
   ```

5.**Running the Action Server** :

The action server was started with the following command to ensure that the actions were properly linked to the chatbot:

```
rasa run actions
```

6.Deployment and Testing:

Once the training was successful, the chatbot was deployed locally for real-time interaction. The deployment command used was:

```
rasa run
```

This command allowed the chatbot to be tested in the terminal. Further testing was conducted to evaluate the chatbot's performance and verify that it responded accurately to user queries

7. SomaBot Deployment:

SomaBot was deployed using Flask, providing a web-based interface for users to interact with the chatbot. Additionally, a website was built to enhance accessibility and user experience.

By following these steps, the chatbot was successfully updated. The modification improved data organization and contributed to the chatbot's maintainability and scalability. Additional refinements and optimizations may be conducted in the future to enhance functionality and improve user experience.

Conclusion

In conclusion, the most common inquiries surrounding CBC implementation emphasize self-expression, agricultural education, rural support, creativity, civic responsibility, innovation, teacher support, extracurricular activities, ethical decision-making, and financial literacy. These concerns reflect the key areas stakeholders consider crucial for the curriculum's success and effectiveness.

The NLP-based sentiment analysis using VADER and TextBlob provided valuable insights into public perception of the education system, highlighting key sentiments that can inform strategic decision-making and policy improvements.

The analysis suggests that most people have a positive outlook on CBC, as positive sentiment has consistently led, peaking in 2024. While there are increasing neutral discussions and some mild concerns, the overall trend indicates that the majority of public sentiment is favorable.

The sentiment classification model that performed best was XGBoost, achieving a weighted F1-score of 0.65, indicating strong overall performance in sentiment prediction.

The Rasa model for the somabot demonstrates outstanding performance, achieving an F1 score of 95.7% and 96.1% intent classification accuracy. With high precision and recall, it consistently makes correct predictions, with minimal misclassifications occurring in ambiguous cases.

Recommendation

Strengthen CBC key focus areas: Self-expression, agricultural education, rural support, creativity, civic responsibility, innovation, teacher support, extracurricular activities, ethical decision-making, and financial literacy by developing targeted policies and resources to enhance curriculum effectiveness.

Policymakers should sustain and build on the positive sentiment peak in 2024 by addressing concerns in neutral and negative discussions through public awareness campaigns that clarify misconceptions and reinforce CBC's benefits.

As neutral discussions rise, further analysis is needed to understand shifting public discourse and potential concerns, while implementing feedback mechanisms to continuously improve CBC based on evolving public opinion.

Enhance sentiment classification accuracy by incorporating additional training data, and exploring advanced NLP techniques like transformer-based models.

Enhance chatbot capabilities by training it on more data to improve real-time assistance, curriculum guidance, and automated responses to CBC-related queries.