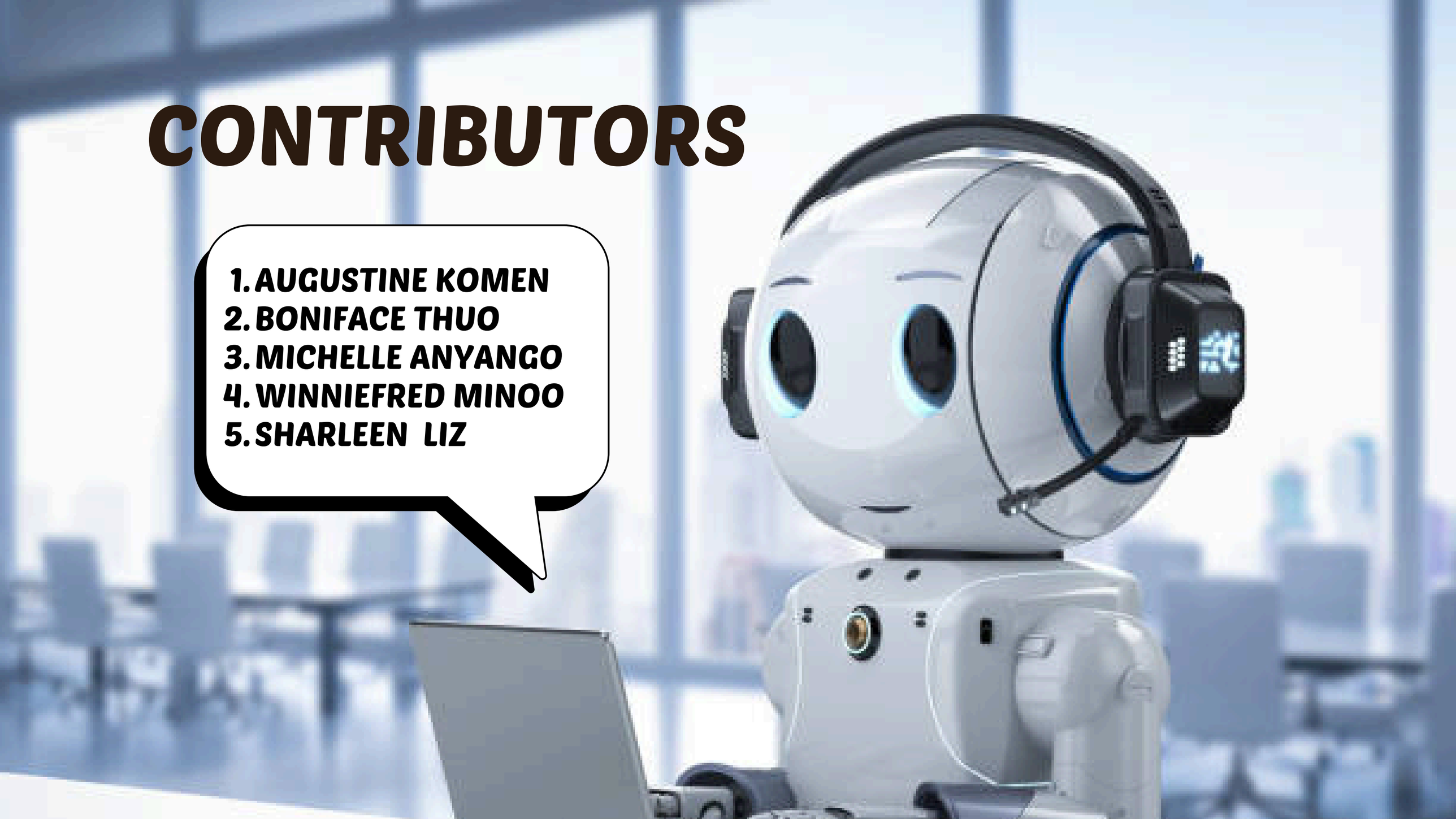
SomaBot

PHASE 5 PROJECT

CONTRIBUTORS

1. AUGUSTINE KOMEN
2. BONIFACE THUO
3. MICHELLE ANYANGO
4. WINNIEFRED MINOO
5. SHARLEEN LIZ



INTRODUCTION

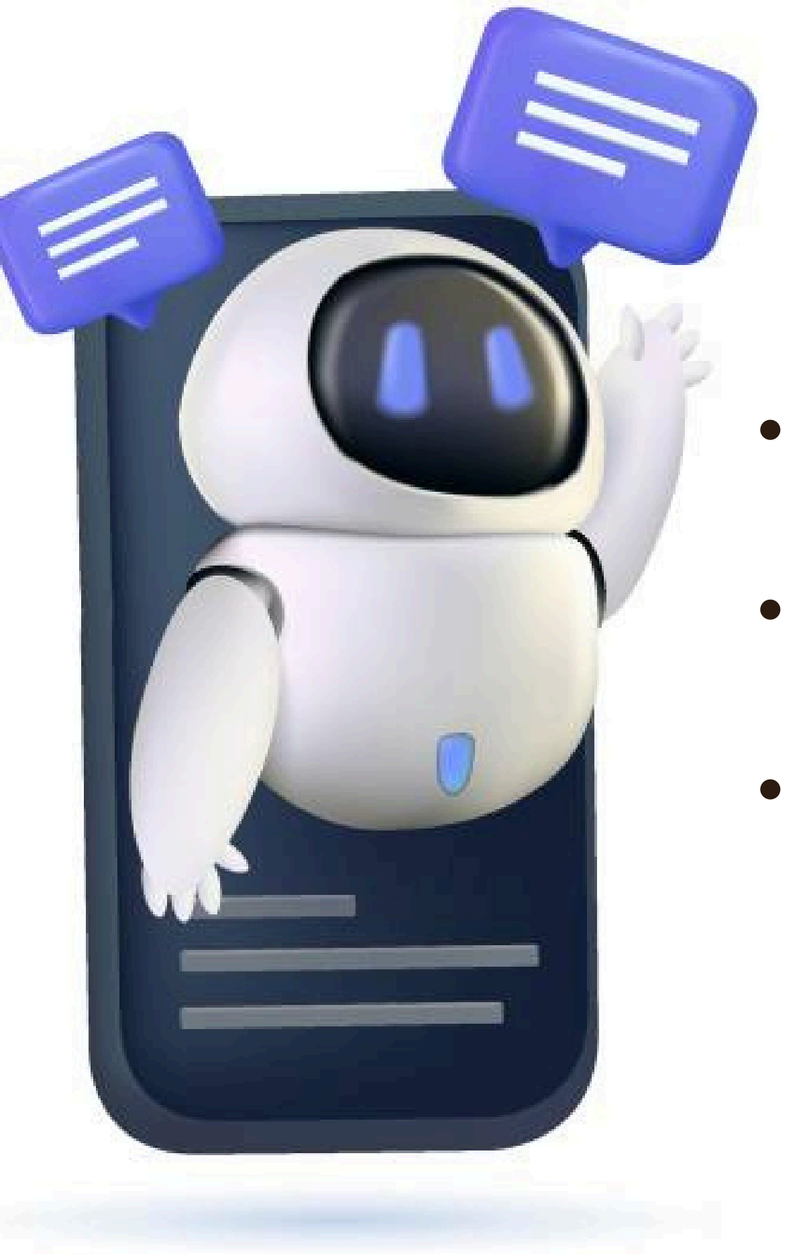
- Chatbots are advanced programs designed to simulate human-like conversations, and their adoption is rapidly transforming industries worldwide.
- SomaBot is a chatbot designed specifically for Kenya's Competency-Based Curriculum (CBC).
- SomaBot will offer instant, accurate information on CBC-related details, assessments, and policies.
- Natural language processing and machine learning will help deliver intelligent, context-aware responses, transforming the way CBC-related information is accessed and understood.

MAIN OBJECTIVE

Develop a chatbot using TF-IDF, Cosine Similarity, and RASA to match user queries with relevant responses, optimizing accuracy and efficiency for precise query resolution.

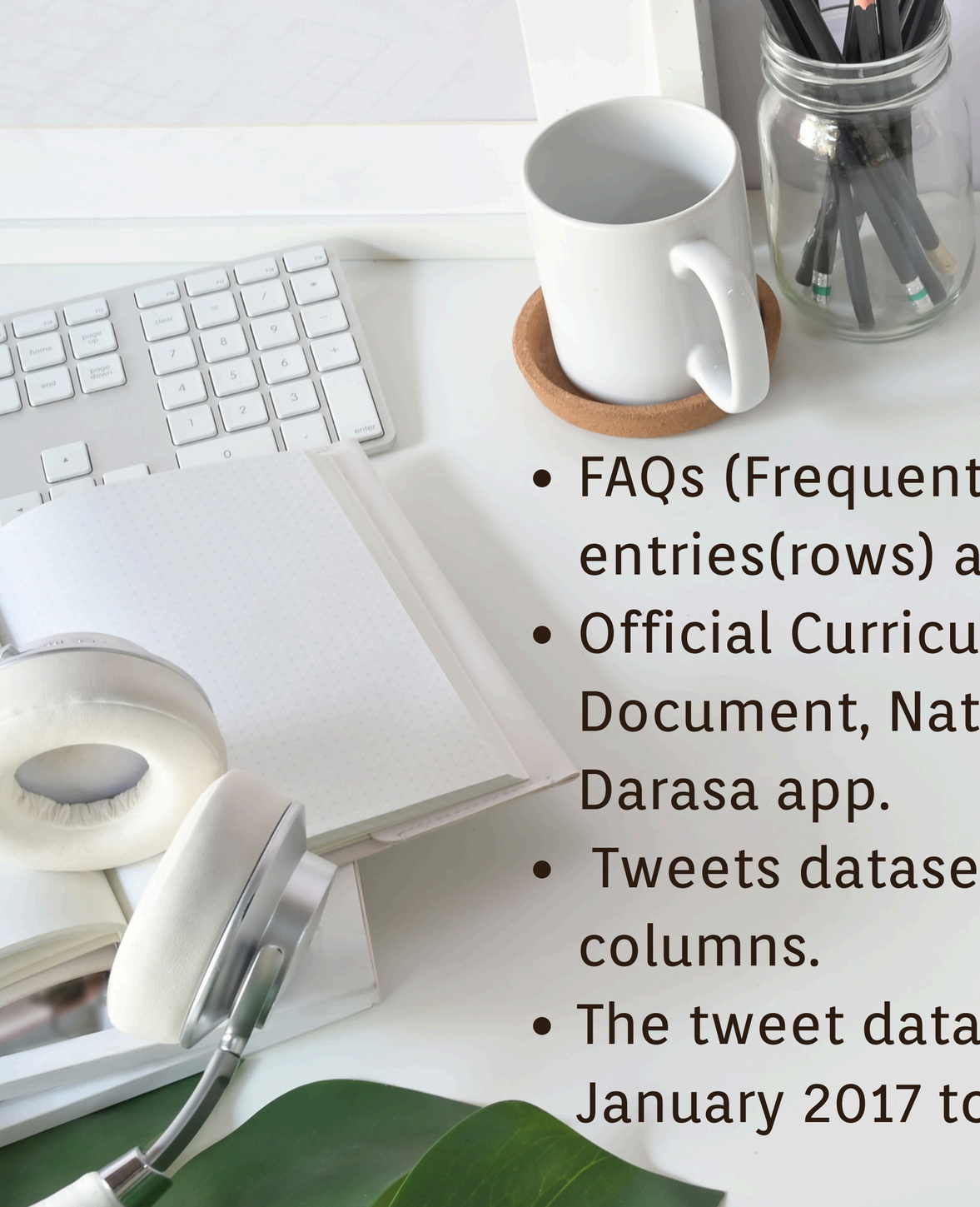
OTHER OBJECTIVES

1. Analyze the FAQS dataset to identify frequently asked queries and common keywords.
2. To apply an NLP-based sentiment analysis approach using VADER and TextBlob to classify CBC-related tweets.
3. To analyze sentiment trends in tweets about the CBC education system.
4. To build and evaluate predictive models for sentiment classification.



SUCCESS METRICS

- The ideal sentiment classification model should attain a minimum F1 score of 65%.
- The top-performing chatbot model should achieve at least 90%.
- To successfully classify tweets into different sentiments (Neutral, Positive, Negative



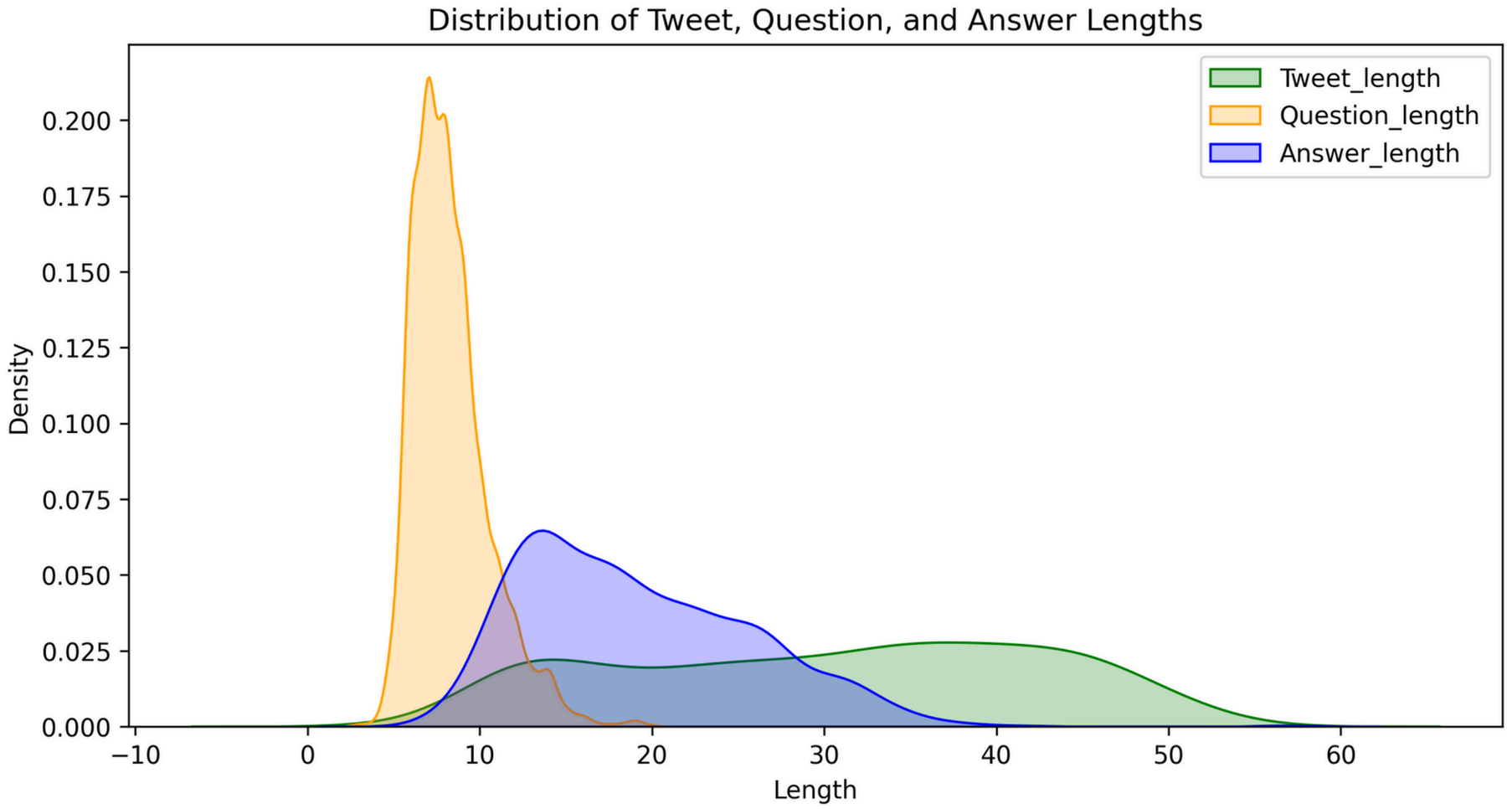
DATA UNDERSTANDING

- FAQs (Frequently Asked Questions) dataset 2533 entries(rows) and 2 (columns).
- Official Curriculum Documents –KICD, Teachers Arena Document, National Curriculum Policy, KNEC, Teacher ac, Darasa app.
- Tweets dataset scrapped tweets 1086 rows and 6 columns.
- The tweet dataset contained CBC tweets from the January 2017 to February 2025.

DATA PREPARATION

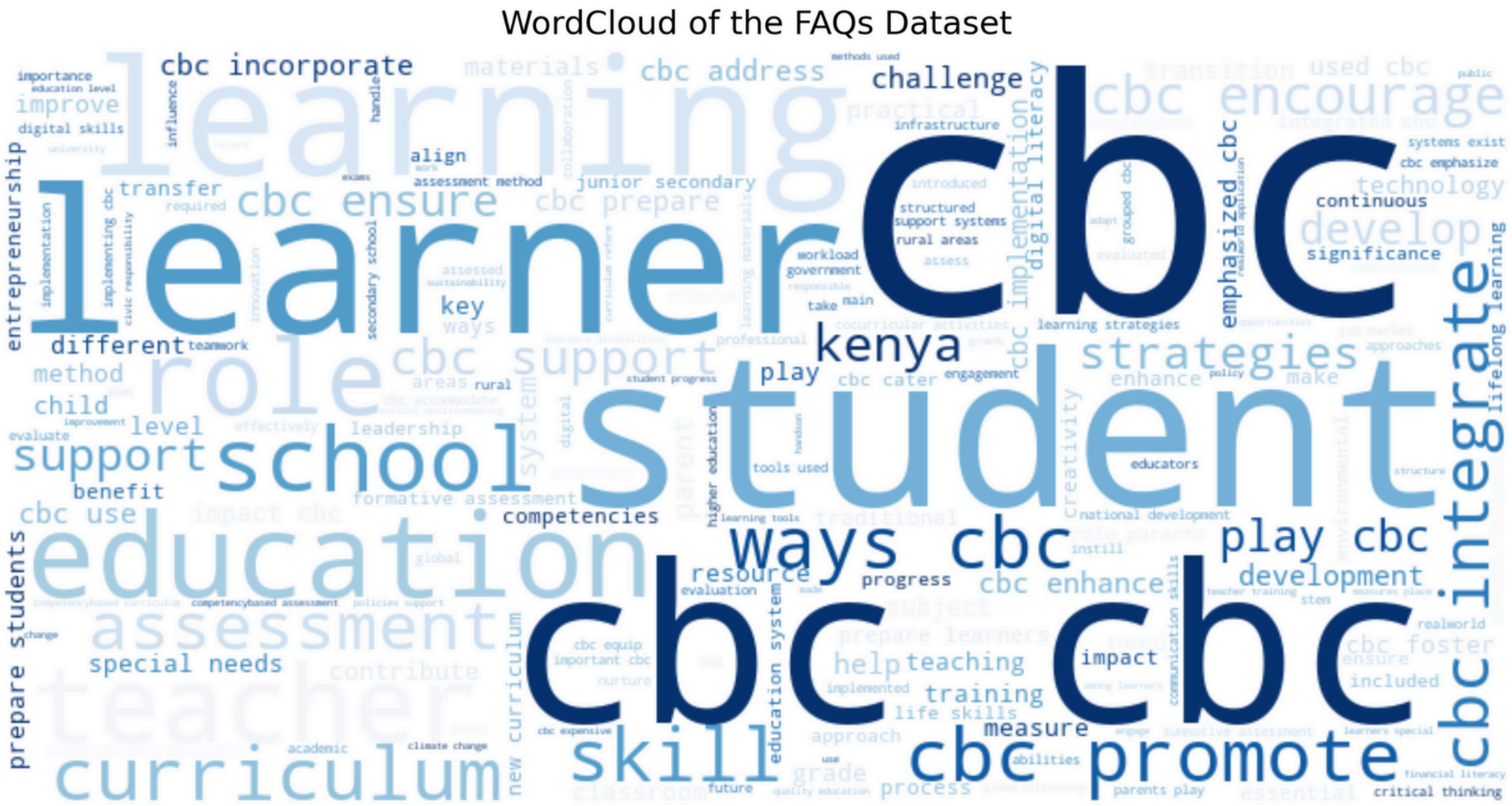
- Dealing with duplicates, outliers and nulls.
- Text cleaning techniques such as removing stopwords.
- Tokenization, lemmatization, and stemming .
- TF-IDF vectorization was used to convert text into numerical representations.
- Dealing with Class imbalance in the tweets dataset

Distribution of tweets ,question and answer lengths



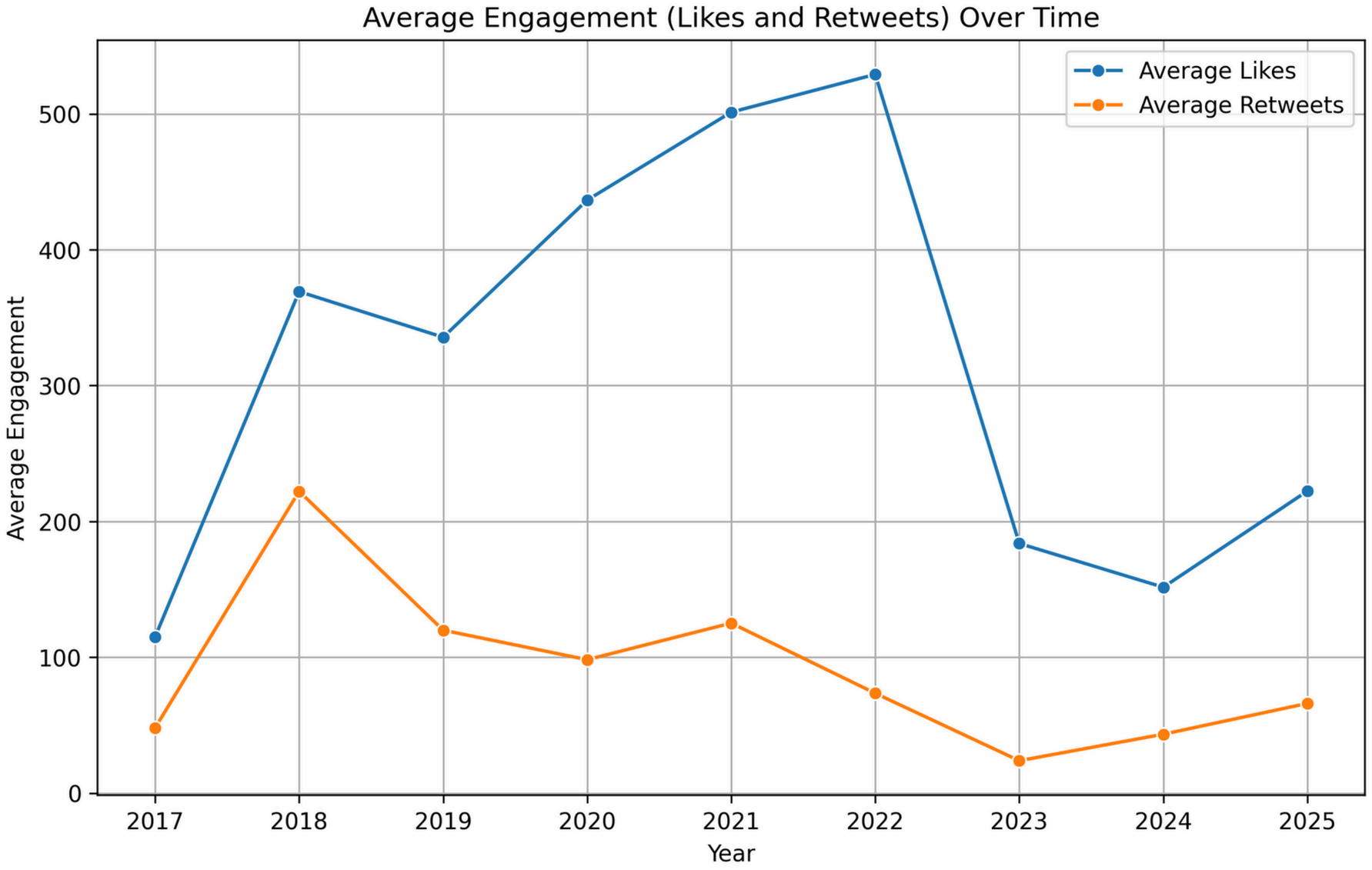
- Question Length: Most questions are brief, peaking below 10 words for clarity and precision.
- Answer Length: Answers range from 10–20 words, adjusting to question complexity.
- Tweet Length: Tweets vary widely, reflecting both concise and detailed expressions.

Common keywords in FAQs dataset



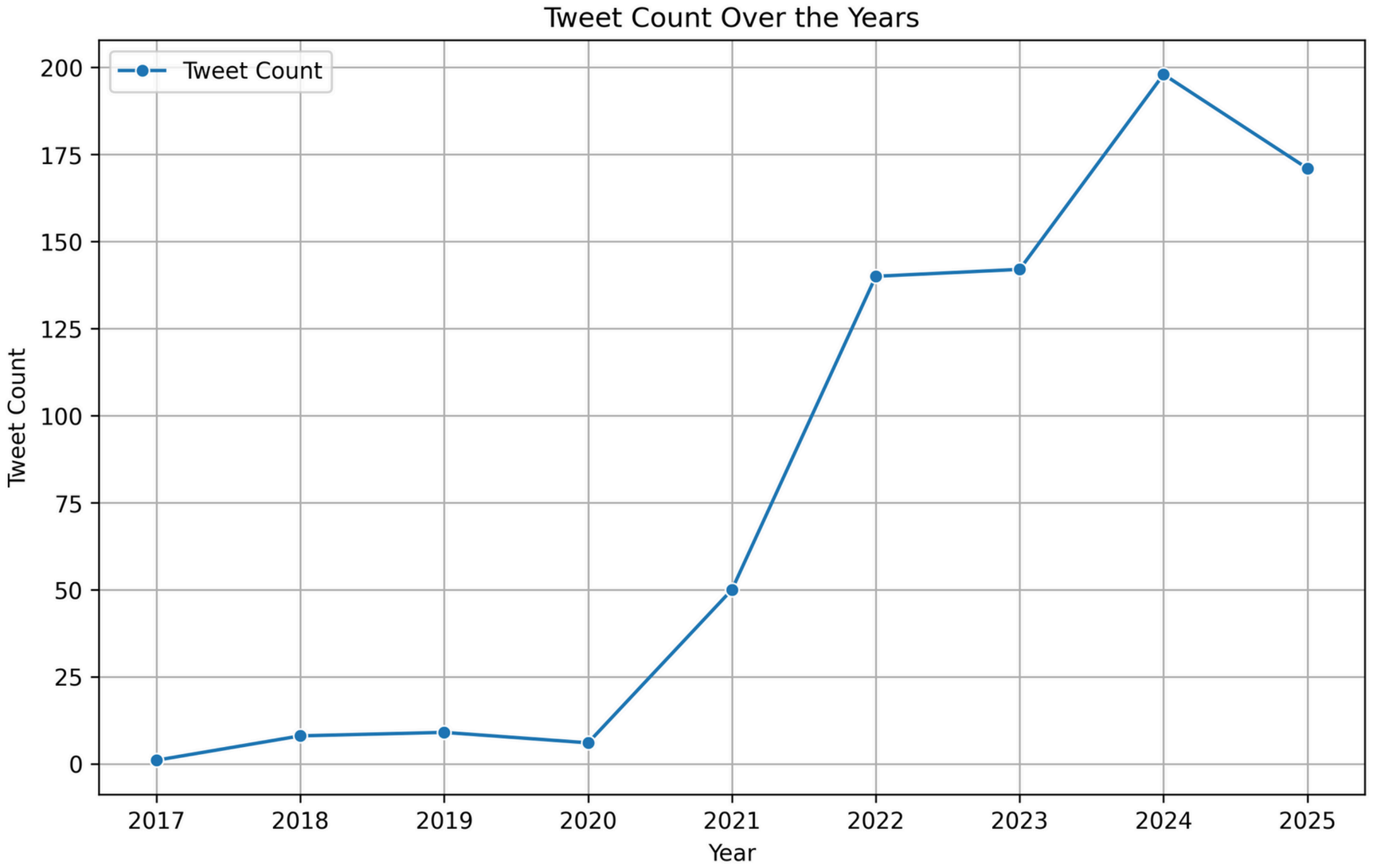
The word cloud highlights discussions on students, educators, and learning, with key terms like "CBC," "learner," and "teacher."

Average engagement (likes and Retweets) overtime



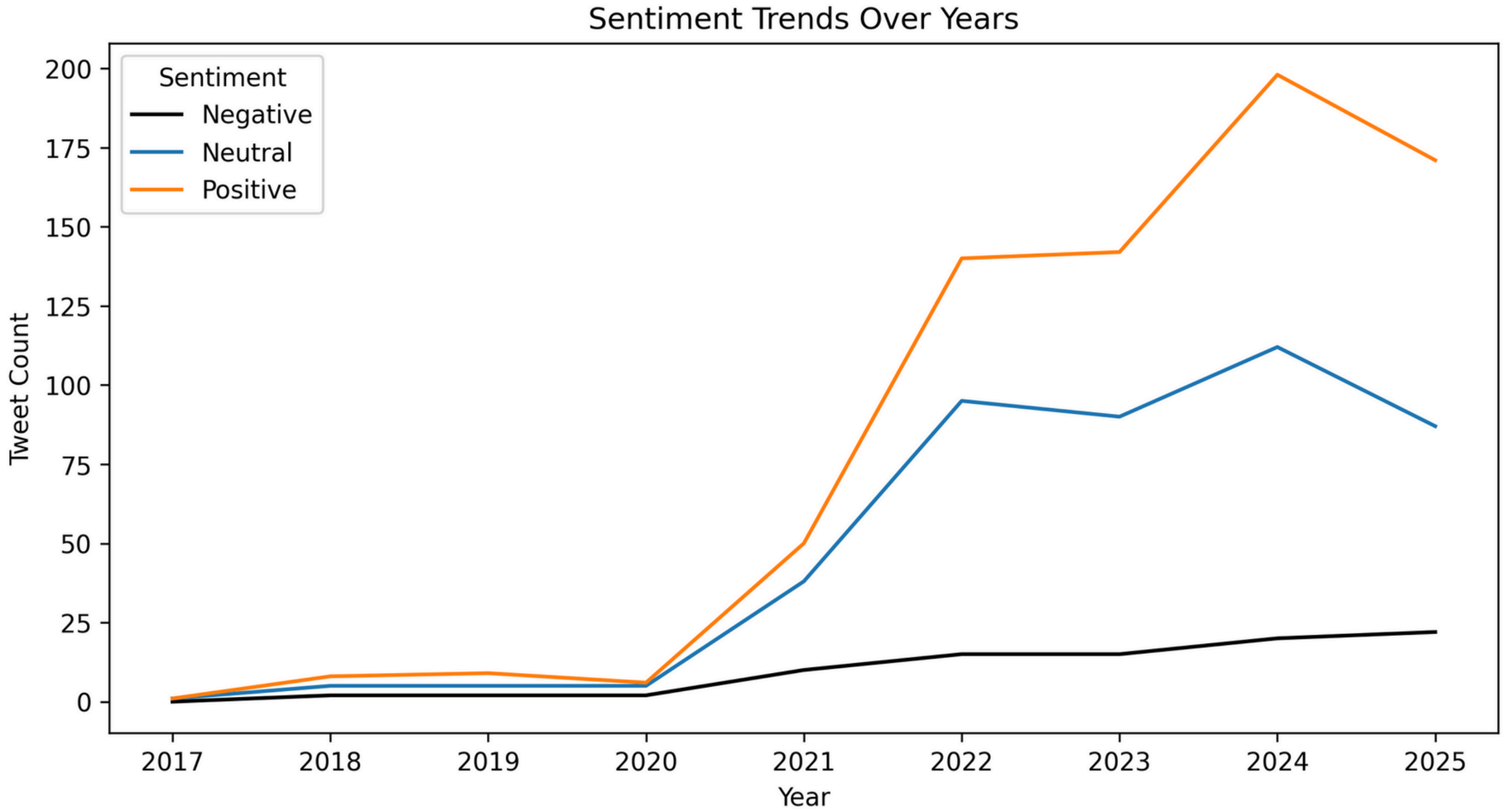
Likes peaked around 2022, followed by a decline, while retweets peaked in 2018 and decreased significantly.

Tweet Count Over the years



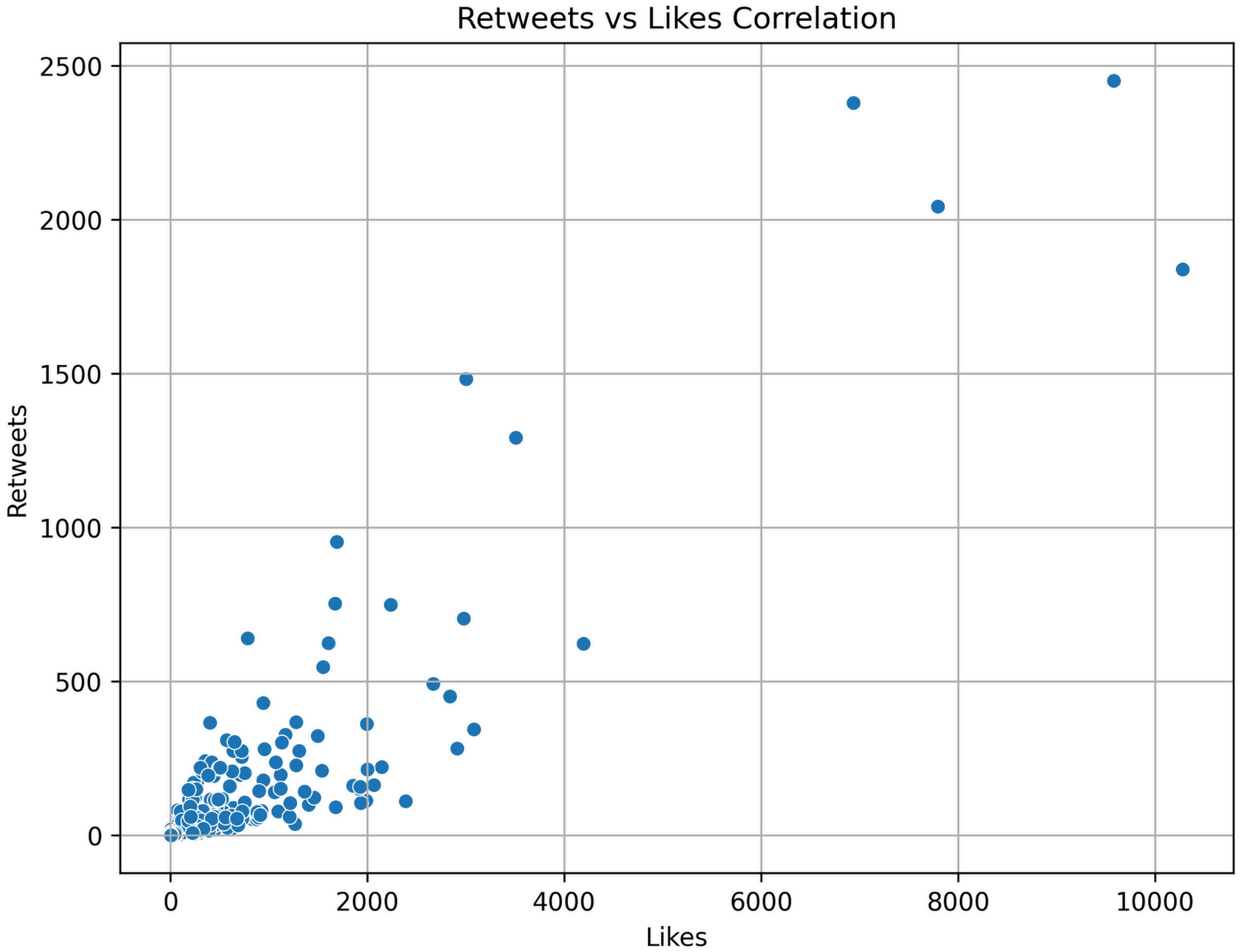
- Tweet activity remained low until 2020, then surged significantly, peaking in 2024.
- A slight decline is observed in 2025, but overall, engagement remains high.

Sentiments Trends Over years

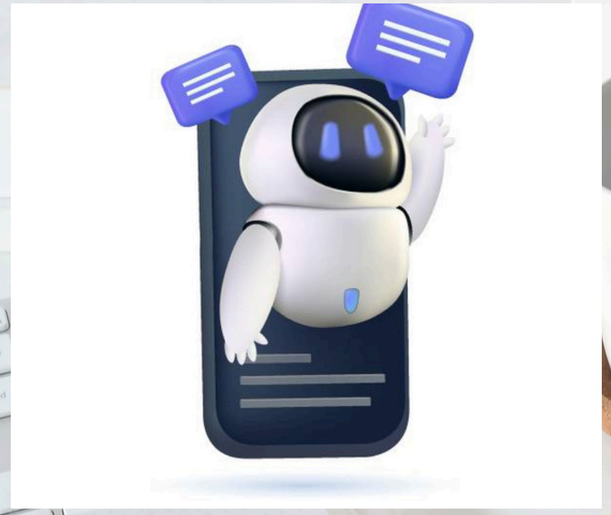


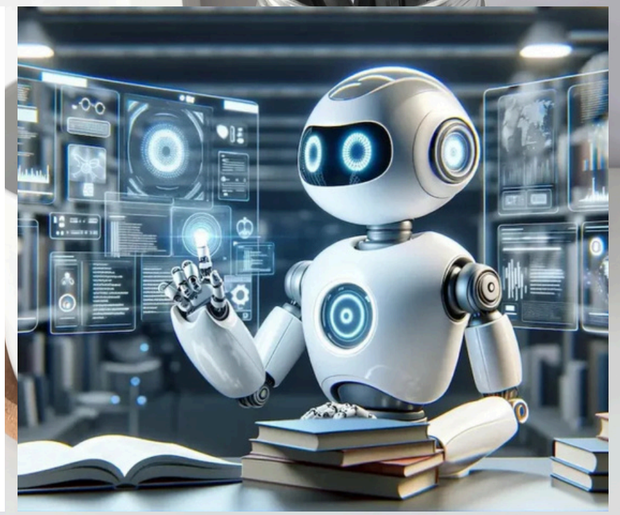
- Tweet activity remained low until 2020, then surged significantly, peaking in 2024.
- A slight decline is observed in 2025, but overall, engagement remains high.

Relationship between retweets and likes



The graph represents a strong correlation between likes and retweets, indicating that highly retweeted tweets also tend to receive more likes, showcasing a strong engagement link.

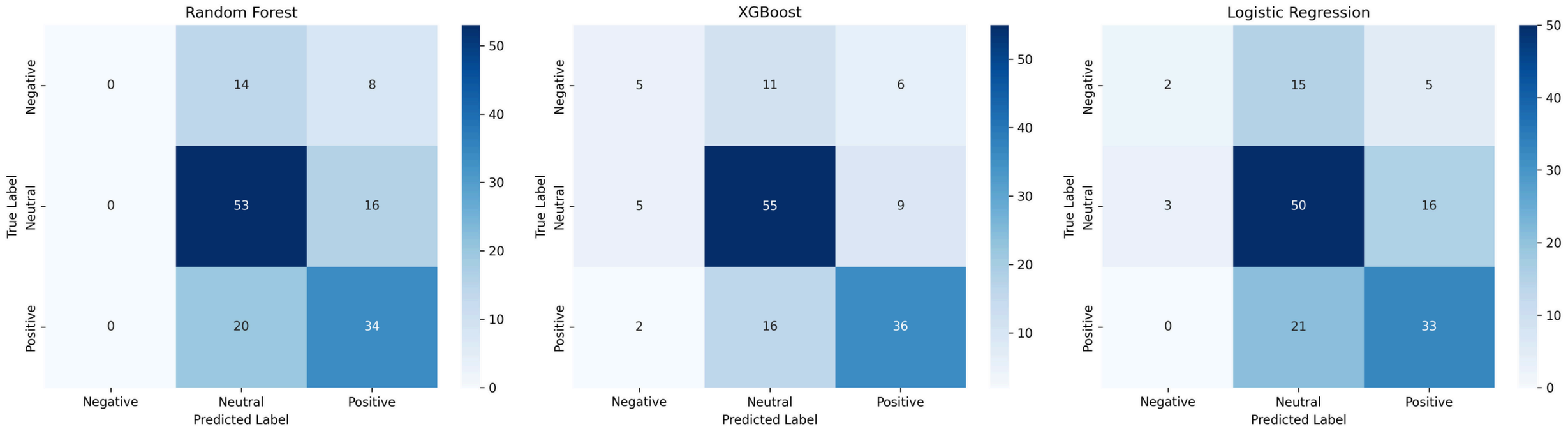




MODELING

- Sentiment Analysis: Random Forest, XGBoost, Logistic Regression, and Neural Networks were incorporated.
- For the CBC-related tweet sentiments, with XGBoost it achieved 63% accuracy and the highest F1-score of 66%.
- Chatbot: TF-IDF with Cosine Similarity,
- RASA model was used to enhance the Chatbot achieving F1 score of 95.7% and 96.1% intent classification accuracy.

Evaluating the sentiment classification model :confusion matrix



- Random Forest misclassifies most Negative cases as Neutral or Positive but performs well on Neutral and Positive labels.
- XGBoost shows better balance but still misclassifies some Negative cases as Neutral or Positive.
- Logistic Regression struggles with Negative cases, often predicting them as Neutral, and confuses some Neutral cases with Positive.

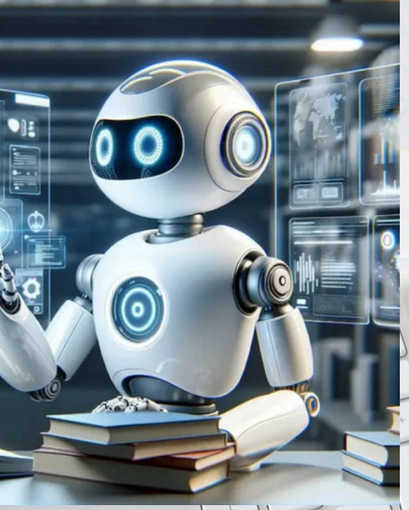
Evaluating the chatbot models TF-IDF

TF-IDF

The TF-IDF chatbot model performed poorly, with only 1% accuracy, precision, recall, and F1-score, indicating severe misclassification issues.

RASA

The evaluation of the Rasa NLU and Core models achieved F1 score of 95.7% and 96.1% intent classification accuracy making it the chatbot's best-performing model.



MODEL DEPLOYMENT

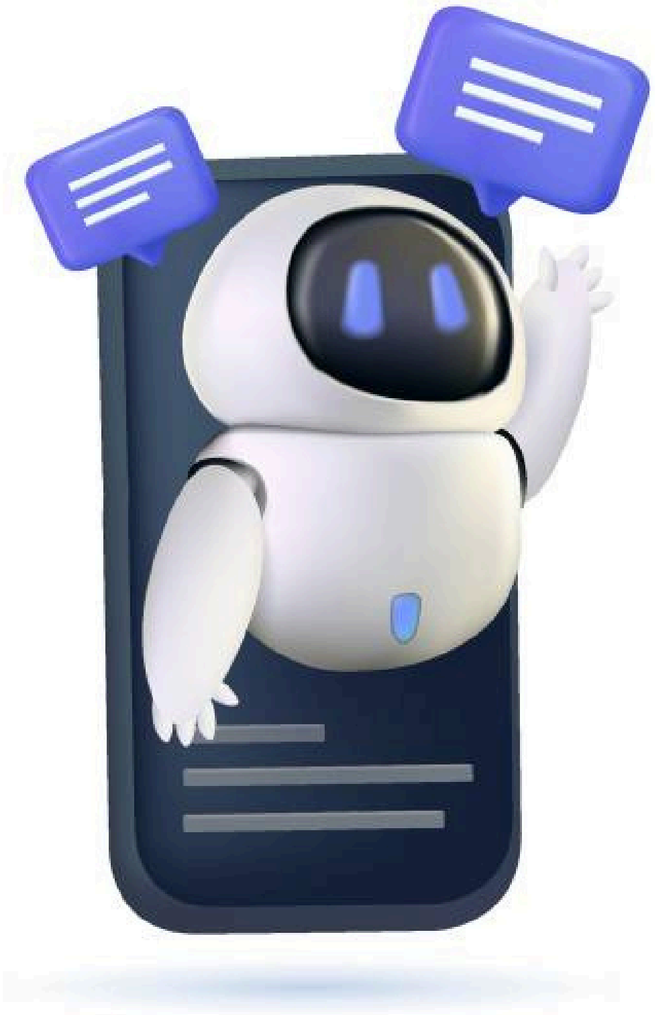
- SomaBot was deployed to an interactive web-based platform.
- It ensured users can easily access curriculum-related information.

CONCLUSION

1. CBC implementation inquiries focus on key areas like self-expression, rural support, and financial literacy.
2. NLP sentiment analysis with VADER and TextBlob reveals public perception trends, guiding policy and strategic decisions.
3. Most people view CBC positively, with sentiment peaking in 2024. Neutral discussions are rising, but overall sentiment remains favorable.
4. The sentiment classification model that performed best was XGBoost, achieving a weighted F1-score of 0.65.
5. The Rasa model excels with a 95.7% F1 score and 96.1% intent accuracy, ensuring precise predictions with minimal errors.

RECOMMENDATIONS

- Enhance CBC by strengthening key areas through targeted policies and resources for improved curriculum effectiveness.
- Policymakers should sustain CBC’s momentum by addressing concerns through awareness campaigns.
- Analyze rising neutral discussions and implement feedback to refine CBC based on public opinion.
- Improve sentiment accuracy with more training data and advanced NLP models like transformers.
- Train the chatbot on more data to improve real-time assistance



THANK YOU

QUESTIONS?