**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 leverages TF-IDF and Cosine Similarity to match user queries with relevant responses, while utilizing Random Forest/XGBoost to rank 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 forstudents, 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**

**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) will serve as the primary data source, supplemented by Ministry of Education policies, circulars, and guidelines. The data will primarily be text-based, comprising 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.

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
* 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)
* APIs – Using the Twitter API or official education portals.
* Manual Curation – Collecting FAQs from government documents.

**Data preparation**

This step ensures that the collected data is clean, structured, and ready for model training.

✔ **Preprocessing Steps:**

* Text Cleaning – Removing special characters, links, and unnecessary whitespace.
* Tokenization – Splitting sentences into words for processing.
* Lemmatization/Stemming – Converting words to their root forms (e.g., *learning → learn*).
* Vectorization – Converting text into numerical format.

**Modelling approach**

**Data preprocessing**

**Conclusion**

**Recommendation**

**References**