1. Business Understanding

Overview
In Kenya's digital landscape, Twitter plays a critical role in shaping public discourse, especially during election seasons and politically charged events. While it provides a platform for civic engagement, it also facilitates the spread of hate speech, tribalism, and inciting language. This content can escalate tensions and contribute to real-world conflict if left unchecked.

Challenges
Manually moderating large volumes of tweets is inefficient and prone to bias. Current moderation systems struggle to scale and often miss localized forms of hate speech, particularly those specific to Kenyan cultural and political contexts. There is also a lack of automated tools specifically designed to detect offensive content within Kenyan tweets.

Proposed Solution
Develop a machine learning model that can automatically categorize tweets into three classes: **hate speech**, **offensive content**, or **neither**. The model will serve as a content flagging tool for:

- Early detection of harmful tweets during elections or national events.
- Support for content moderation teams at media outlets, regulatory bodies, and social platforms.
- Enhancing digital safety by tracking the evolution and spread of dangerous speech trends.

Brief Conclusion
An effective classification model tailored to the Kenyan context can reduce the spread of inciting content and improve the quality of online public discourse. This system would be vital in promoting peace, especially during sensitive periods like elections.

Problem Statement

There is an urgent need for a localized and scalable hate speech detection system to flag inciting and offensive tweets in Kenya, particularly during political periods where online speech can incite real-world unrest.

Objectives

- To evaluate the most common terms or phrases used in inciting tweets.

- To build a multi-class classification model that labels tweets as **hate speech**, **offensive**, or **neither**.
- To evaluate model performance across different algorithms using F1 score.
- To Deploy the best performing model using Streamlit

---

2 Data Understanding

The dataset used in this project, **HateSpeech_Kenya.csv**, was sourced from **Kaggle**, a well-known platform for sharing datasets and machine learning challenges. It contains approximately **48,000 tweets** labeled for sentiment and hate speech, specifically focused on the Kenyan social and political context.

Key Features of the Data:

- **tweet**: The raw text of the tweet . this forms the primary input for natural language processing and classification tasks.
- **label**: The target variable, where 0 represents **neutral** content and 1 represents **offensive** content and 2 represent **hate speech**.

This dataset is directly aligned with the project goal of developing a machine learning model to detect and categorize hate speech in Kenyan tweets. The inclusion of localized language, real-world offensive content, and metadata allows the model to learn patterns specific to the Kenyan context.

3. Data Preparation

Initial Observations:

- The tweets contain a variety of **noisy text elements**, including **URLs, emojis, hashtags, user mentions, and inconsistent punctuation**.
- The language is primarily **English**, but many tweets feature **code-switching with Swahili** and **Kenyan slang**, which may impact language modeling.
- A noticeable **class imbalance** exists, with a higher proportion of neutral tweets compared to hate/offensive ones.

Planned Modifications and Processing Steps:

1. **Text Cleaning and Normalization**
   The tweets will undergo standard preprocessing to make them suitable for modeling, including:
   - Lowercasing text
   - Removing URLs, mentions, hashtags, emojis, numbers, and punctuation
   - Tokenization followed by **lemmatization** or **stemming**
   - Removing stopwords and **Kenyan-specific filler words**
2. **Feature Engineering**
   To enhance model input, we will extract and generate relevant features:
   - Representing text using **TF-IDF** and exploring **word embeddings**
   - Creating numerical features such as **tweet length**, **word count**, and **sentence count**
3. **Handling Class Imbalance**
   Given the skew in class distribution, we will use the following method to mitigate imbalance:
   - **Undersampling** of the majority class

These planned steps will ensure that the data is clean, structured, and ready for building a robust hate speech detection model tailored to the Kenyan digital context.

4. Modeling Approach

The project will apply **supervised machine learning techniques** to classify tweets as **hate/offensive/neutral** based on the textual content of each post. The goal is to develop an accurate and generalizable classifier capable of detecting harmful speech patterns in the Kenyan social media context.

Primary models under consideration include:

- **Traditional models** such as **Logistic Regression**, **Naïve Bayes** (well-suited for text classification tasks), and **ensemble methods** like **Random Forest** and **XGBoost** for efficient training and strong baseline performance.
- **Transformer-based deep learning models**, such as BERT.

The modeling approach will involve a comparative analysis between traditional machine learning algorithms and advanced deep learning architectures designed for natural language understanding.

5 Evaluation

The models will be thoroughly evaluated to identify the best-performing one for deployment. Evaluation will focus primarily on **F1 score**. Additional metrics such as **ROC-AUC** and the **confusion matrix** will be used to assess overall classification performance and understand the distribution of correct and incorrect predictions across all classes.

6. Deployment Plan

The best-performing model will be deployed as an interactive web application using Streamlit, a lightweight Python framework ideal for rapid prototyping and user-friendly interfaces.

Key Deployment Steps:

- Model Integration: The trained and validated model will be saved using joblib or pickle, and loaded into the Streamlit app for real-time predictions.
- User Input: Users will be able to input tweet text manually
- Prediction Output: The app will return the predicted class
- Hosting: The app will be hosted on Streamlit Cloud

Deployment Objective:
To provide an accessible tool for moderators, journalists, and civil society to detect and flag harmful tweets, especially during politically sensitive periods in Kenya.