DATA DOCUMENTATION REPORT - THE INK PULSE PROJECT

The data used for this project was collected from a single primary source. The source was an Excel file called The Hill Speaks (Responses). xlsx, which contained raw survey responses written partly in Ghanaian Pidgin English. These responses were collected through a Google Form and then downloaded into Excel. After that, the Excel file was processed to combine the weekly responses into one text column. The result was saved as a CSV file called weekly_raw_text_sequential.csv. This CSV file made it easier to work with the data during preprocessing and analysis.

Before the data could be analysed, several cleaning and preparation steps were carried out. First, the raw Pidgin English text was translated into Standard English using the Gemini API. A prompt carefully guided the translation to ensure the original text's cultural meaning and informal style were still kept. A retry system was added to handle any possible translation failures. The translated data was saved in a csv file called weekly_translated_text(2).csv which is our cleaned dataset.

After translating the text, it was cleaned by removing emojis using regular expressions. Punctuation marks like commas, periods, and question marks were also removed. All words were then converted into lowercase letters to make the text uniform. Next, the text was split into words using a regular expression tokeniser. Common English words that do not add much meaning, such as "the," "is," and "at," were removed. Then, a lemmatisation step was applied, changing words to their root form, such as "running" to "run." Any missing values found in the data were also handled by converting them into strings to avoid errors during processing.

To better understand the data, we created some simple visualisations. We plotted a histogram to show the number of words each week, making it easy to observe the activity level in weekly responses. A Word Cloud was also used to display the most common words creatively. Additionally, we created a bar chart to highlight the top 10 most frequently mentioned words, which reflected the main topics discussed during the survey period.

We employed descriptive statistics to analyse the distribution of response lengths from students. A histogram was generated to illustrate the spread of sentiment polarity scores, helping us understand the range of emotions expressed by the students. Furthermore, we included a line graph to show the frequency of weekly postings, indicating how often students submitted their concerns or shared their thoughts.

These visualizations provide crucial insights that inform the decision to use topic modelling algorithms like LDA and BERTopic. The variability in weekly posting frequency and peak activity during evenings suggests temporal patterns in student engagement that topic models can track over time. The word cloud and sentiment distribution highlight recurring themes and emotional tones such as stress and dissatisfaction, indicating distinct, underlying topics suited for extraction. Then, the variation in response lengths points to the need for robust models like BERTopic, which can handle short and detailed texts more effectively than traditional methods, justifying the combined use of LDA and BERTopic to capture the breadth and depth of student concerns.