XEBIA PROJECT REPORT

**Week 1 Report: Sentiment Analysis, Importing Libraries, Data Collection, and Preprocessing**

**Enrolment:21CS002364**

**Name: Biradhar Nithin Kumar  
Branch: CSE (AI & ML) [Batch 2]**

**Mentor : Sevy**

**1. Introduction**

In the first week, the focus was on setting up the environment, collecting the data, and performing initial preprocessing to prepare the dataset for analysis and modeling.

**2. Importing Libraries**

The following libraries were imported to handle various tasks:

* **warnings**: To ignore warning messages and keep the output clean.
* **pandas**: For data manipulation and analysis.
* **seaborn and matplotlib.pyplot**: For data visualization.
* **nltk**: For natural language processing, including downloading and using stopwords.
* **emoji**: For handling emojis in the text data.
* **re**: For using regular expressions to process text data.
* **WordCloud and Counter**: For creating word clouds and counting word frequencies.
* **sklearn**: For building and evaluating machine learning models.

**3. Data Collection**

Two CSV files containing Twitter data were loaded into pandas DataFrames:

* twitter\_training.csv
* twitter\_validation.csv

These datasets were concatenated into a single DataFrame, final\_df, which included columns for tweet IDs, entities, sentiment targets, and tweet content.

**4. Data Preprocessing**

**Checking for Missing Values:**

* The code checked for missing values in the dataset. The content column had some missing values.
* Rows with missing values were removed using dropna().

**Handling Duplicates:**

* The code checked for duplicate rows and removed them using drop\_duplicates(). This ensured that the dataset was clean and free of redundant entries.

**Feature Engineering:**

* New features were created to provide additional insights into the text data:
  + **Character Count**: The number of characters in each tweet.
  + **Word Count**: The number of words in each tweet.
  + **Sentence Count**: The number of sentences in each tweet.

These features were computed using pandas and NLTK functions.

**Text Preprocessing:**

* The text data was converted to lowercase to maintain consistency.
* Punctuation was removed using string.punctuation.
* Stopwords were removed using NLTK's list of English stopwords.
* Words were stemmed using the PorterStemmer to reduce words to their root form.

**Handling Emojis:**

* Emojis were removed from the text data using the emoji library, ensuring that only textual content was processed.

**Week 2 Report: Exploratory Data Analysis (EDA)**

**1. Introduction**

The second week was focused on performing exploratory data analysis (EDA) to understand the dataset better and gain insights into the distribution and characteristics of the data.

**2. Target Variable Distribution**

**Visualizing Sentiment Distribution:**

* A pie chart was created to visualize the proportion of each sentiment category (Positive, Negative, Neutral, Irrelevant) in the dataset.
* A bar plot was generated to show the count of each sentiment category, providing a clear view of the sentiment distribution.

**3. Feature Distribution**

**Character, Word, and Sentence Count Distribution:**

* Distributions of the newly created features (character count, word count, sentence count) were visualized using dist plots. This helped in understanding the spread and central tendency of these features.

**4. Text Preprocessing Analysis**

**Word Clouds:**

* Word clouds were generated for each sentiment category to visualize the most frequent words in Positive, Negative, Neutral, and Irrelevant tweets. These visualizations provided a quick overview of the common words associated with each sentiment.

**Most Common Words:**

* The code identified the 50 most common words for each sentiment category and plotted them using bar plots. This analysis helped in understanding the key terms associated with different sentiments.

**Week 3 Report: Model Building, Evaluation, and Summary**

**1. Introduction**

The third week focused on building machine learning models, evaluating their performance, and summarizing the results.

**2. Label Encoding**

**Converting Sentiment Labels to Numerical Values:**

* Sentiment labels were encoded into numerical values to facilitate model training:
  + Positive: 1
  + Negative: 0
  + Neutral: 2
  + Irrelevant: 3

**3. Data Splitting**

**Splitting Data into Training and Testing Sets:**

* The dataset was split into training and testing sets with an 80-20 split. This ensured that the models were trained on a substantial portion of the data while being evaluated on unseen data.

**4. Model Building**

**RandomForestClassifier:**

* A pipeline was created using TfidfVectorizer for text vectorization and RandomForestClassifier for classification.
* The model was trained on the training data and evaluated on the testing data.

**MultinomialNB:**

* Another pipeline was created using TfidfVectorizer and MultinomialNB.
* This model was also trained and evaluated similarly.

**LogisticRegression:**

* A third pipeline was created using TfidfVectorizer and LogisticRegression.
* The model was trained and evaluated, providing an additional comparison.

**5. Model Evaluation**

**Performance Metrics:**

* The accuracy score of each model was calculated and displayed.
* Confusion matrices were generated for each model, providing a detailed view of the model's performance in classifying different sentiment categories.

**6. Model Saving**

**Saving the Best Model:**

* The RandomForestClassifier model, which performed the best among the three models, was saved using pickle for future use.

**Summary**

* **Week 1**: Focused on setting up the environment, collecting data, and performing initial preprocessing to clean and prepare the dataset.
* **Week 2**: Involved performing EDA to gain insights into the data distribution and characteristics, including visualizing sentiment distribution and common words.
* **Week 3**: Concentrated on building and evaluating multiple machine learning models, comparing their performance, and saving the best model for future use.

**Week 4 Task Report: Dashboard Creation for sentiment Analysis:**

**Objective:**

The objective was to create an interactive dashboard in Power BI to visualize the sentiment analysis results. This included generating various charts and plots to present the data in a comprehensive manner.

**Importing CSV into Power BI**

* Data Import: Imported the CSV file into Power BI to create visual representations of the sentiment analysis data.

