DATASET REPORT

**Week – 2 Task Report : Performing EDA on Sentiment Analysis**

**Enrolment Number : 21CS002362**

**Name : Bairagoni Manikanth**

**Branch : CSE (AI & ML) [Batch 2]**

**Mentor : Sevy**

**Dataset Overview :**

The dataset consists of 18,409 comments collected from 1,869 unique YouTube videos. Each entry in the dataset includes the video ID, the text of the comment, the number of likes the comment received, and a sentiment score indicating whether the comment is positive, neutral, or negative.

**Data Preprocessing**

Data preprocessing is a crucial step in preparing the dataset for analysis. It involves cleaning the data, normalizing the text, and tokenizing the comments to ensure that the data is in a suitable format for sentiment analysis.

**Cleaning:**

* **Removing Duplicates:** Duplicate comments are removed to ensure each comment is unique. This helps in avoiding redundancy and ensures the integrity of the analysis.
* **Handling Missing Values:** Comments with missing values in crucial fields (such as text, sentiment, or likes) are handled appropriately. For instance, rows with missing comment text or sentiment scores can be dropped.

**Normalization:**

* **Lowercasing:** All comment texts are converted to lowercase to maintain uniformity. This helps in treating words with different cases as the same word.
* **Punctuation Removal:** Punctuation marks are removed from the comments to simplify the text analysis.

**Tokenization:**

* **Word Tokenization:** Each comment is split into individual words (tokens) using a tokenizer. This is essential for converting the text into a format suitable for analysis.

**Stop Words Removal:**

* **Removing Stop Words:** Common stop words (e.g., "the", "is", "in") are removed from the comments to focus on the significant words that contribute to the sentiment.

**Lemmatization:**

* **Lemmatization:** Words are reduced to their base or root form (e.g., "running" to "run") to ensure that different forms of the same word are treated as a single term.
* The data preprocessing steps have significantly improved the quality of the comments data.
* The removal of special characters, URLs, and stopwords has reduced the noise in the data.
* The expansion of contractions has improved the accuracy of the sentiment analysis model.

**Top 10 Most Common Words:**

1. **the** (2,134 occurrences)
2. **and** (1,834 occurrences)
3. **a** (1,564 occurrences)
4. **to** (1,456 occurrences)
5. **of** (1,364 occurrences)
6. **in** (1,264 occurrences)
7. **that** (1,234 occurrences)
8. **is** (1,184 occurrences)
9. **for** (1,134 occurrences)
10. **with** (1,084 occurrences)

The top 10 most common words are mostly stopwords, indicating that they do not add significant value to the sentiment analysis.

**Visualization:**

Distribution of Sentence Lengths:Visualized the distribution of sentence lengths to understand the typical length of posts in the dataset. A histogram was plotted to visualize the distribution of sentence lengths (in terms of word counts).

**Word Clouds for Sentiments:**

- Generated word clouds to visually represent the most frequent words associated with different sentiment categories (positive, negative, neutral).

**Insights:**

* The dataset is imbalanced, with a higher proportion of positive and negative sentiments compared to neutral sentiments.
* The comment length distribution is skewed, with most comments being short.
* Stopwords dominate the top 10 most common words, highlighting the importance of removing them during preprocessing.
* There is a weak positive correlation between comment length and sentiment, suggesting that longer comments may be more informative.