Week 2 Task Report: Performing EDA on Sentiment Analysis Project

**Enrolment:21CS002377**

**Name: Gudapelly Shrihari**

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

**1. Dataset Overview:**

- The dataset consists of social media posts with associated metadata such as timestamp, platform, and sentiment.

- It spans various countries and includes posts from different users across multiple platforms like Twitter, Instagram, and Facebook.

**2. Preprocessing:**

Text Cleaning:All text data was converted to lowercase and special characters were removed to standardize the text.

Tokenization and Stopword Removal: Text was tokenized into words, and common stopwords were removed to focus on meaningful content.

Lemmatization: Words were lemmatized to reduce inflectional forms and normalize the text for analysis.

Handling Missing Values: Rows with missing values were dropped to ensure data integrity. Removing Duplicates: Duplicate rows based on `username`, `comment`, and `date` were removed to avoid bias in analysis.

**3. Analysis Techniques Applied:**

Sentence Type Classification: Implemented to categorize sentences into declarative, interrogative, exclamatory, and imperative types based on sentence structure.

**4.POS Tagging:**

Parts of speech tagging was applied to identify the grammatical components of each word in the processed text.

**5.Named Entity Recognition (NER):**

Identified and categorized named entities such as persons, organizations, locations, etc., using SpaCy's NER capabilities.

**6.Vowel Counting:**

Calculated the number of vowels in each post to explore linguistic characteristics.

**7. Sentiment Analysis:**

Used VADER (Valence Aware Dictionary and sEntiment Reasoner) to compute sentiment scores (positive, negative, neutral, compound) for each post.

**8.Dependency Parsing:**

Analyzed the syntactic structure of sentences to understand relationships between words (subject, object, modifiers, etc.).

**9.Topic Modeling:**

Applied Latent Dirichlet Allocation (LDA) to discover underlying topics within the posts, revealing common themes across the dataset.

**10.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).

**11.Word Clouds for Sentiments:**

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

**12.Insights:**

Popular Topics:Topics such as travel, fitness, movies, and cooking were prevalent based on the identified topics.

Sentiment Trends: Overall, a majority of posts expressed positive sentiments, followed by neutral and negative sentiments.

User Engagement: Posts related to nature, fitness, and gratitude received higher engagement (likes and retweets), indicating popular topics among users.

**Conclusion**

This report summarizes the comprehensive analysis of the sentiment dataset, showcasing the application of various NLP techniques to derive meaningful insights from social media posts.