**Name:Angothu Muni**

**Enrollement Number:21CS002361**

**College:Spsu**

**Group: Batch 1**

**Analysis Report on IMDB Movie Reviews Dataset**

**1. Introduction**

The IMDB Movie Reviews dataset consists of 1000 randomly sampled reviews. This report summarizes the preprocessing steps, exploratory data analysis (EDA), and various natural language processing (NLP) techniques applied to extract meaningful insights from the dataset.

**2. Data Preprocessing**

* **Null Values and Duplicates:**
  + Null values were checked across the dataset. No null values were found.
  + Duplicate reviews were identified and removed, resulting in 0 duplicates.

**3. Text Cleaning**

* **Remove HTML Tags:**
  + HTML tags within the 'review' column were cleaned using BeautifulSoup, ensuring text purity for subsequent analysis.

**4. Word Count Analysis**

* **With and Without Stopwords:**
  + The number of words in each review was computed both including and excluding stopwords (common words like 'the', 'is', etc.).

**5. Distribution of Sentence Lengths**

* **Histogram of Word Counts:**
  + A histogram was plotted to visualize the distribution of sentence lengths (in terms of word counts).

**6. N-gram Analysis**

* **Bi-grams and Tri-grams:**
  + The most common bi-grams and tri-grams were extracted to identify frequent word pairs and triplets.

**7. Sentence Types**

* **Classification of Sentence Types:**
  + Sentences were categorized as declarative, exclamatory, or interrogative based on the presence of punctuation marks.

**8. Part-of-Speech (POS) Tagging**

* **POS Tagging:**
  + Each review was tagged with its part-of-speech (POS) to identify the syntactic role of each word.

**9. Named Entity Recognition (NER)**

* **NER Analysis:**
  + Named entities such as persons, organizations, and locations were identified using NLTK's named entity chunker.

**10. Count of Vowels**

* **Vowel Count:**
  + The total number of vowels in each review was calculated to explore linguistic characteristics.

**11. Sentiment Analysis**

* **Sentiment Scores:**
  + Polarity and subjectivity scores were computed using TextBlob to gauge the sentiment expressed in each review.

**12. Syntactic Analysis**

* **Dependency Parsing:**
  + Dependency parsing was performed using spaCy to analyze the grammatical structure of sentences.

**13. Word Cloud Visualization**

* **Word Cloud:**
  + A word cloud was generated to visualize the most frequent words in the reviews, providing a quick overview of the most discussed topics.

**14. Topic Modeling**

* **Latent Dirichlet Allocation (LDA):**
  + Topic modeling using LDA identified latent topics within the reviews, showcasing the underlying themes present in the dataset.

**Conclusion**

This report encapsulates a comprehensive analysis of the IMDB Movie Reviews dataset, leveraging various NLP techniques to uncover insights about review characteristics, sentiment, syntactic structure, and topical focus. The application of these techniques not only enhances our understanding of the dataset but also highlights the utility of NLP in extracting meaningful information from textual data.