Report

Train Dataset

1. Vocabulary and Word Characteristics

- **Vocabulary Size:** 111,839 unique words.
- **Average Word Length:** 5.45 characters.

The dataset boasts a rich vocabulary, which indicates a wide variety of terms and expressions used. The average word length of 5.45 characters suggests a balanced mix of shorter and longer words, typical for a diverse and comprehensive text corpus.

2. Part-of-Speech (POS) Distribution

The distribution of parts of speech within the dataset is as follows:

- **Nouns (NOUN):** 3,863,775
- **Proper Nouns (PROPN):** 2,395,417
- **Verbs (VERB):** 2,168,497
- **Adjectives (ADJ):** 1,009,447
- **Adverbs (ADV):** 329,376
- **Punctuation (PUNCT):** 240,086
- **Interjections (INTJ):** 120,552
- **Adpositions (ADP):** 87,810
- **Numerals (NUM):** 56,901
- **Auxiliary Verbs (AUX):** 41,765
- **Spaces (SPACE):** 20,349
- **Pronouns (PRON):** 14,881
- **Particles (PART):** 11,965
- **Subordinating Conjunctions (SCONJ):** 10,552
- **Other (X):** 7,752
- **Determiners (DET):** 1,658
- **Symbols (SYM):** 1,631
- **Coordinating Conjunctions (CCONJ):** 1,604

The prevalence of nouns and proper nouns suggests that the dataset is rich in subjects and entities. Verbs and adjectives are also well-represented, supporting diverse syntactic structures and descriptive content.

3. Sentence Structure

- **Average Sentence Length:** 35.91 words.

The average sentence length is relatively high, indicating complex and informative sentences.

4. Named Entity Recognition (NER) Distribution

The entity distribution within the dataset includes:

- **PERSON:** 367,468

- **ORG (Organizations):** 82,348
- **DATE:** 47,856
- **GPE (Geopolitical Entities):** 47,145
- **CARDINAL:** 38,573
- **TIME:** 36,460
- **NORP (Nationalities, Religions, Political Groups):** 27,542
- **ORDINAL:** 9,968
- **PRODUCT:** 4,574
- **QUANTITY:** 4,529
- **LOC (Locations):** 3,725
- **FAC (Facilities):** 3,629
- **MONEY:** 2,643
- **LANGUAGE:** 995
- **EVENT:** 818
- **WORK_OF_ART:** 699
- **LAW:** 522
- **PERCENT:** 339

This distribution highlights a strong presence of personal names, organizations, and temporal entities, indicating that the dataset contains a significant amount of biographical or news-related content.

5. Sentiment Analysis

- **Average Sentiment Polarity:** 0.03
- **Average Sentiment Subjectivity:** 0.43

The near-neutral average sentiment polarity (0.03) suggests that the dataset is balanced in terms of positive and negative sentiments. The subjectivity score of 0.43 indicates that the content includes a mix of objective and subjective statements.

6. Scene Analysis

- **Total Scenes:** 824
- **Average Scene Length:** 142.60
- **Max Scene Length:** 518
- **Min Scene Length:** 3
- **Label Distribution:** Counter({0.0: 100,825, 1.0: 16,678})
- **Number of Unique Tokens:** 382,112
- **Average Number of Labels 0 per Scene:** 122.36
- **Average Number of Labels 1 per Scene:** 20.24

The dataset is divided into 824 scenes, with an average length of 142.60, indicating substantial segments of text. The maximum scene length of 518 and minimum of 3 words show a wide variance in scene length, accommodating both detailed and concise scenes. The label distribution with a higher prevalence of 0.0 labels compared to 1.0 labels indicates an imbalanced dataset, which is a crucial consideration for training models.

The dataset contains a large number of unique tokens, signifying a diverse lexicon. The average number of labels per scene further emphasizes the imbalance towards the 0.0 label.

Test Dataset

Total scenes: 50

Average scene length: 163.12

Max scene length: 646

Min scene length: 44

Label distribution: Counter({0.0: 5468, 1.0: 2688})

Number of unique tokens: 78930

Average number of labels 0 per scene: 109.36

Average number of labels 1 per scene: 53.76

Vocabulary and Word Characteristics

- **Vocabulary Size:** 32,138 unique words.
- **Average Word Length:** 5.43 characters.

2. Part-of-Speech (POS) Distribution

The distribution of parts of speech within the dataset is as follows:

- **Nouns (NOUN):** 238,532
- **Proper Nouns (PROPN):** 147,379
- **Verbs (VERB):** 132,937
- **Adjectives (ADJ):** 64,716
- **Adverbs (ADV):** 19,059
- **Punctuation (PUNCT):** 16,438
- **Interjections (INTJ):** 7,171
- **Adpositions (ADP):** 5,247
- **Numerals (NUM):** 4,813
- **Auxiliary Verbs (AUX):** 2,593
- **Spaces (SPACE):** 1,280
- **Pronouns (PRON):** 776
- **Particles (PART):** 664
- **Subordinating Conjunctions (SCONJ):** 621
- **Other (X):** 603
- **Coordinating Conjunctions (CCONJ):** 137
- **Determiners (DET):** 99
- **Symbols (SYM):** 53

3. Sentence Structure

- **Average Sentence Length:** 33.97 words.

4. Named Entity Recognition (NER) Distribution

The entity distribution within the dataset includes:

- **PERSON:** 23,290
- **ORG (Organizations):** 5,807
- **DATE:** 3,319
- **CARDINAL:** 3,137
- **GPE (Geopolitical Entities):** 2,487
- **TIME:** 2,272
- **NORP (Nationalities, Religions, Political Groups):** 1,590
- **ORDINAL:** 749
- **PRODUCT:** 692
- **QUANTITY:** 347
- **FAC (Facilities):** 264
- **LOC (Locations):** 141
- **MONEY:** 132
- **LANGUAGE:** 86
- **LAW:** 39
- **WORK_OF_ART:** 37
- **EVENT:** 34
- **PERCENT:** 23

5. Sentiment Analysis

- **Average Sentiment Polarity:** 0.04
- **Average Sentiment Subjectivity:** 0.42

6. Gender Word Distribution

The gender word distribution within the dataset includes:

- **Male:** 2,844
- **Female:** 835

This distribution shows a higher frequency of male-gendered words compared to female-gendered words, indicating a potential gender bias in the content.

Summary

The dataset is extensive and diverse, featuring a wide range of vocabulary, complex sentence structures, and a variety of named entities.