EDA and Preprocessing Report (Updated)

December 27, 2024

Introduction

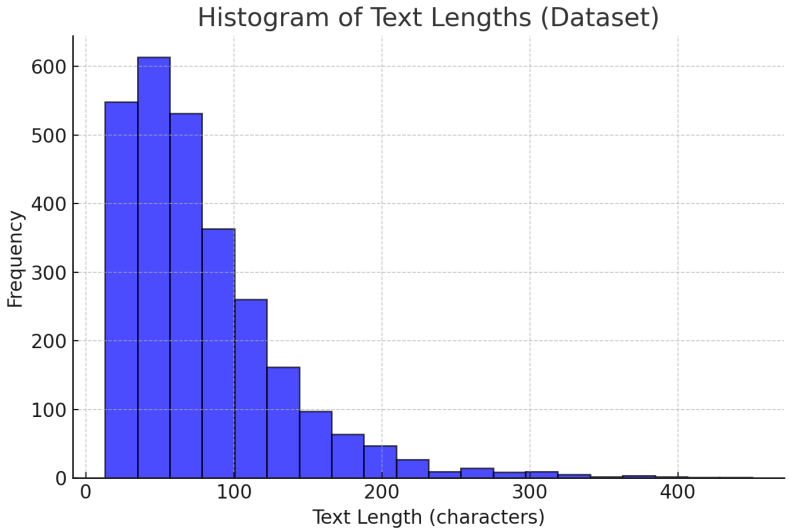
This report outlines the exploratory data analysis (EDA) and preprocessing steps performed on the updated dataset for the SemEval competition project. The goal is to prepare the data for downstream tasks such as emotion classification.

Dataset Overview

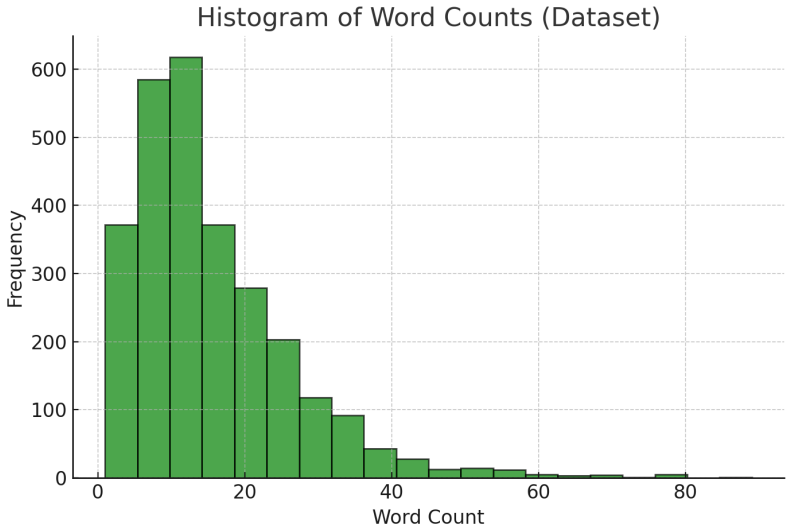
- **Number of Rows:** 2768
- **Number of Features:** 7
- **Missing Values:** None
- **Class Distribution:** Imbalanced, with specific distributions shown below.

Distribution Analysis

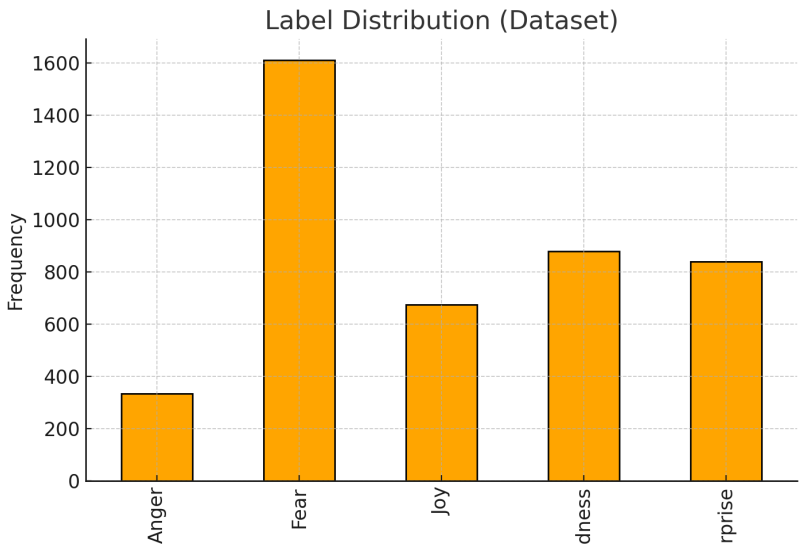
Text Length Distribution



Word Count Distribution



Label Distribution



Sample Preprocessed Data (Training Dataset)

Below is a sample of the preprocessed training dataset:

ID	Original Text	Preprocessed Text
eng_train_track_a_00001	But not very happy.	but not very happy
eng_train_track_a_00002	Well she's not gon na last the whole song like that.	well shes not gon na last the whole song like that
eng_train_track_a_00003	She sat at her Papa's recliner sofa only to move.	she sat at her papas recliner sofa only to move
eng_train_track_a_00004	Yes, the Oklahoma city bombing.	yes the oklahoma city bombing
eng_train_track_a_00005	They were dancing to Bolero.	they were dancing to bolero

Preprocessing Steps

The following steps were applied to preprocess the text data:

1. Lowercasing: All text converted to lowercase.
2. Removing Punctuation and Numbers: Stripped all punctuation and numeric characters.
3. Tokenization: Split text into individual words.
4. Stopword Removal: Removed common stopwords.

5. Lemmatization/Stemming: Words reduced to their root forms using lemmatization or stemming.

References

- NLTK Tokenizer Documentation
- Semantic Analysis Overview
- Text Data Preparation Guide