Natural Language Processing Project

AMAZON PRODUCT REVIEWS SENTIMENT ANALYSIS

Abstract:
This NLP project aims to perform sentiment analysis on Amazon US reviews specifically for Mobile Electronics products using the DistilBERT language model. Sentiment analysis is the process of determining the emotional tone behind a piece of text, in this case, customer reviews. By employing DistilBERT, a powerful transformer-based language model, we seek to extract valuable insights from the reviews to understand customers' sentiments towards various mobile electronics products available on Amazon US. The results of this analysis will help businesses and consumers gain a better understanding of customer feedback and make informed decisions based on the sentiment expressed in these reviews.

Methodology:

1. Data Collection: We will collect a large dataset of Amazon US reviews for Mobile Electronics products. The dataset will include various attributes such as product ratings, review body(text), and other relevant metadata.
2. Data Preprocessing: The collected data will undergo preprocessing steps, including text cleaning, tokenization, and removal of stop words and special characters. Additionally, we will handle any missing or erroneous data points to ensure the quality and integrity of the dataset.
3. DistilBERT Model: We will utilize the DistilBERT language model, a lightweight version of BERT, to conduct the sentiment analysis. DistilBERT offers similar performance to BERT while being computationally more efficient, making it suitable for this project.
4. Fine-Tuning: To make the DistilBERT model task-specific, we will fine-tune it on our Mobile Electronics reviews dataset. This process involves adjusting the model's weights using the labeled sentiment data to optimize its performance for sentiment analysis.
5. Training and Evaluation: We will split the dataset into training and testing sets to train the fine-tuned DistilBERT model. The model's performance will be evaluated on the testing set using metrics to assess its effectiveness in sentiment classification.
6. Sentiment Analysis: Once the model is trained and evaluated, we will apply it to new, unseen reviews for Mobile Electronics products to predict their sentiment. The sentiment will be classified into positive or negative categories.

Technical Terms:

1. Sentiment Analysis: The process of determining the sentiment expressed in a piece of text, i.e., whether it is positive or negative.

2. Transformer: A transformer is a deep learning architecture that utilizes self-attention mechanisms to process sequential data efficiently, revolutionizing various natural language processing tasks.
3. Hugging Face: A popular organization that develops and maintains a repository of pre-trained language models, including DistilBERT, making it accessible and easy for everyone to use NLP models for various tasks.
4. DistilBERT: A lightweight, distilled version of the BERT (Bidirectional Encoder Representations from Transformers) language model, known for its efficiency and competitive performance in natural language processing tasks.
5. Tokenization: The process of breaking down text into smaller units called tokens, which can be words, subwords, or characters, to facilitate language model processing.