Sentiment Analysis Report

Geat Aharon

Description of the Dataset Used

The dataset named "amazon_product_reviews.csv" contains Amazon product reviews that have different attributes such as dates, ID numbers, URLs, manufacturer details, product ratings, and review texts. For sentiment analysis, the main focus is on the 'reviews.rating' and 'reviews.text' columns, which provide numerical ratings and textual feedback from users, respectively. Other data points such as dates, ID numbers, URLs, and manufacturer details are considered irrelevant for this analysis.

Preprocessing Steps involved in the analysis

- Dataset Loading: The "amazon_product_reviews.csv" is loaded with `low_memory=False` to handle large datasets efficiently in a Pandas DataFrame.
- Sentiment Mapping: Ratings from the 'reviews.rating' column are mapped to 'Positive', 'Neutral', or 'Negative' based on their value, to facilitate comparison with the model's predictions.
- Missing Values Removal: Rows missing 'reviews.text' are dropped using `dropna()`, ensuring the analysis only includes complete reviews.
- Sentiment Analysis Application: The 'analyze_sentiment' function classifies each review's sentiment using spaCy and SpacyTextBlob, based on text polarity.
- DataFrame Copy: A copy of the DataFrame is made with `.copy()` to avoid `SettingWithCopyWarning` during sentiment analysis.

Evaluation of Results

The sentiment analysis model showcased an accuracy of 0.8245, indicating that over 82% of the sentiment predictions matched the original sentiment derived from the numerical ratings. Additionally, the model tested the semantic similarity between two randomly selected reviews and assessed sentiment for a sample of five product reviews, demonstrating its ability to interpret and analyze textual sentiment effectively.

Insights into the Model's Strengths and Limitations

Strengths:

- Simplicity and Efficiency: The use of spaCy and SpacyTextBlob provides a straightforward and efficient approach to sentiment analysis.
- Good Performance: With an accuracy rate of 82.45%, the model demonstrates a strong capability in sentiment classification.
- Versatility in Analysis: The ability to analyze sentiment and compute similarity between texts offers diverse insights.

Limitations:

- Fixed Polarity Thresholds: Reliance on static thresholds for sentiment categorization may not always capture nuanced emotional expressions accurately.
- Model's Generalization: Performance might vary significantly with texts differing in domain or style from the training data.
- Dependence on spaCy's Language Model: The model's effectiveness is tied to the underlying spaCy language model, which might not include the latest linguistic features or slang.