Lambton College | In Mississauga

AML 2304 – NATURAL LANGUAGE PROCESSING 01

Project Report

Sentiment Analysis on Amazon Food Review

Date of Submission: December 13th 2023

Instructor: Mr. Bhavik Gandhi

Group Members

Gitik Kaushik	(C0867079)
Harcharan Singh	(C0863764)
Yajur Sethi	(C0863424)
Tega Orido	(C0857045)
Anargha Manoj	(C0865188)

Names	Task Completed
Harcharan Singh	Text processing and sentiment analysis with NLTK, sentiment analysis with transformers(pipeline)
Yajur Sethi	Data visualization with Matplotlib and Seaborn, examples and interpretations,
Gitik Kaushik	Data visualization with seaborn.
Tega Orido	Integration of sentimental analysis.
Anargha Manoj	Data loading & exploration, sentiment analysis with transformers (Roberta).

Index

Project Report: Sentiment Analysis on Amazon Food Review

1. Introduction:

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) task that involves determining the sentiment expressed in a piece of text. In this project, our aim is to conduct sentiment analysis on a dataset of Amazon reviews. We utilize two different approaches: rule-based sentiment analysis with VADER and machine learning-based sentiment analysis with a pre-trained RoBERTa model.

2. Objectives:

Analyze the sentiment of Amazon reviews using both rule-based and machine learning-based methods.
Compare the results obtained from VADER and RoBERTa to understand their respective strengths and limitations.
Visualize sentiment distributions and explore relationships with review scores.

3. Data Loading and Exploration:

We start by loading the dataset (Reviews.csv) using the Pandas library and explore its structure to understand the available attributes.

We provided a code that aids in identifying potential data issues, such as missing values or incorrect data types, allowing for necessary data cleaning or preprocessing steps.

4. Data Preprocessing:

We preprocess the data by selecting a subset of the dataset for analysis. Additionally, we handle any missing or irrelevant data.

This step involves transforming raw data into a more usable format, making it suitable for analysis, modeling, or machine learning tasks

It includes Identifying and dealing with missing or null values in the dataset. Techniques include imputation (replacing missing values), deletion of rows/columns, or using algorithms that can handle missing data

5. Exploratory Data Analysis (EDA):

Visualizing the distribution of review scores provides insights into the overall sentiment of the dataset.

This is a critical initial phase in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset

6. Natural Language Processing (NLP) with NLTK:

We demonstrate basic NLP techniques such as tokenization and part-of-speech tagging using NLTK.

We are using programming techniques and algorithms to analyze, understand, and derive meaning from human language in a computational manner

Key Concepts and Techniques in NLP with NLTK:

a. Tokenization:

- **Definition:** Breaking down text into smaller units, such as sentences or words (tokens).
- **NLTK Functionality:** NLTK provides methods to tokenize text into sentences or words for further analysis.

b. Part-of-Speech (POS) Tagging:

- **Definition:** Assigning grammatical tags (e.g., noun, verb, adjective) to words in a sentence.
- **NLTK Functionality:** NLTK offers tools for POS tagging, which is essential for syntactic and semantic analysis.

c. Named Entity Recognition (NER):

- **Definition:** Identifying and classifying named entities (e.g., names of persons, organizations, locations) in text.
- **NLTK Functionality:** NLTK includes NER modules to extract named entities from text.

d. Sentiment Analysis:

- **Definition:** Determining the sentiment (positive, negative, neutral) expressed in text.
- **NLTK Functionality:** NLTK offers sentiment analysis tools like VADER (Valence Aware Dictionary and sentiment Reasoner) for analyzing sentiments in text.

e. Stemming and Lemmatization:

- **Definition:** Reducing words to their base or root forms to normalize text.
- **NLTK Functionality:** NLTK provides modules for stemming and lemmatization, aiding in text normalization.

7. Sentiment Analysis with VADER:

We utilize VADER, a rule-based sentiment analysis tool from NLTK, to calculate sentiment scores for each review.

Key Features of VADER:

- **Rule-Based Approach:** VADER employs a set of rules and a sentiment lexicon to assign sentiment scores to text based on predetermined thresholds.
- **Valence Scores:** It provides polarity scores (positive, negative, neutral, and compound) for text, indicating the intensity and direction of sentiment

8. Sentiment Analysis with RoBERTa:

We employ a pre-trained RoBERTa model from Hugging Face's Transformers library for machine learning-based sentiment analysis.

Key Features of RoBERTa:

- **Transformer Architecture:** RoBERTa utilizes a transformer architecture, specifically designed to understand and process sequential data like text.
- **Pre-Trained Model:** It comes pre-trained on vast amounts of text data, enabling it to capture intricate language patterns and nuances.
- **Contextual Understanding:** RoBERTa's bidirectional nature enables it to understand contextual relationships between words in a sentence.

Functionality and Use Cases:

- **Sentiment Classification:** RoBERTa can be fine-tuned for sentiment analysis tasks, classifying text into categories like positive, negative, or neutral sentiment.
- **High Performance:** It often achieves state-of-the-art performance in various NLP tasks due to its powerful contextual understanding and fine-tuning capabilities.
- **Adaptability:** RoBERTa's pre-trained nature allows for fine-tuning on specific datasets or domains, enhancing its performance on targeted tasks.

9. Sentiment Analysis on the Entire Dataset:

Sentiment Analysis on the Entire Dataset serves as a comprehensive analysis of sentiments expressed across all samples, enabling a holistic understanding of the collective sentiment trends or distributions. While providing valuable insights, it's essential to consider the tool's limitations and contextual nuances inherent in the dataset to derive accurate and meaningful conclusions.

10. Data Visualization:
Visualize relationships between sentiment scores and review scores using pair plots.

Visual Representations:

- **Charts and Graphs:** Including bar charts, line graphs, histograms, scatter plots, pie charts, heatmaps, etc., each suited for different data types and purposes.
- **Maps and Geospatial Visualization:** Representing data geographically to explore spatial relationships or regional variations.

b. Techniques:

- **Color Coding:** Using colors effectively to encode information, distinguish categories, or highlight specific data points.
- **Interactivity:** Incorporating interactive elements for exploration and deeper analysis

11. Examples of High and Low Sentiment Reviews:

Explore examples of reviews with high and low sentiment scores from both VADER and RoBERTa.

Examples of High and Low Sentiment Reviews refer to specific instances or samples from a dataset that represent extreme sentiments, either very positive (high sentiment) or very negative (low sentiment)

12. Hugging Face's Sentiment Analysis Pipeline:

Demonstrate the simplicity and efficiency of Hugging Face's sentiment analysis pipeline.

Hugging Face's Sentiment Analysis Pipeline is a streamlined and efficient tool that leverages pre-trained transformer models for quick and accurate sentiment analysis. Here's a brief summary:

Objective:

- **Performing Sentiment Analysis:** Offers a straightforward and efficient method to analyze sentiment in text data without extensive manual setup or customization.

13. Sentiment Analysis Result

Review:

cat food i bought was expired.

Sentiment Analysis Result:

VADER Result: {'vader_neg': 0.0, 'vader_neu': 1.0, 'vader_pos': 0.0, 'vader_compound': 0.0, 'roberta_neg': 0.83157873, 'roberta_neu': 0.16091663, 'roberta_pos': 0.007504545}

RoBERTa Result: NEGATIVE (Confidence: 0.997921884059906)

Review:

candies i bought tasted good but later i noticed it was expired

Sentiment Analysis Result:

VADER Result: {'vader_neg': 0.0, 'vader_neu': 0.822, 'vader_pos': 0.178, 'vader_compound': 0.2382, 'roberta_neg': 0.58390796, 'roberta_neu': 0.3494014, 'roberta_pos': 0.06669061}

RoBERTa Result: NEGATIVE (Confidence: 0.9911825656890869)

Review:

this is a very health dog food which dog was happily eating for more than 2 years

Sentiment Analysis Result:

VADER Result: {'vader_neg': 0.0, 'vader_neu': 0.795, 'vader_pos': 0.205, 'vader_compound': 0.5574, 'roberta_neg': 0.0053466256, 'roberta_neu': 0.090787336, 'roberta_pos': 0.90386605}

RoBERTa Result: POSITIVE (Confidence: 0.9976525902748108)

14. Conclusion:

This project provides a comprehensive analysis of sentiment in Amazon reviews using both rule-based (VADER) and machine learning-based (RoBERTa) approaches. The comparison between these methods offers valuable insights. Visualizations enhance the understanding of sentiment distributions and their relationships with review scores.

15. Future Enhancements:

Explore other pre-trained models for sentiment analysis.

Perform sentiment analysis on larger datasets for a broader understanding.
Implement fine-tuning of models on domain-specific data for improved accuracy.

16. Acknowledgments:

We acknowledge the NLTK, Hugging Face Transformers, and other open-source libraries that contributed to the success of this project.

17. References:

NLTK Documentation: https://www.nltk.org/
Hugging Face Transformers Documentation: https://huggingface.co/transformers/

18. Code Availability:

The complete code for this project is available ON
https://github.com/Harcharan1997/NLP-Project