Goldsmiths
UNIVERSITY OF LONDON

Python Sentiment Analysis Project

A Report submitted for Natural Language Processing Module

March 27, 2024

Module leader: Dr Akshi Kumar

Carlos Manuel De Oliveira Alves
Student ID: 33617310

Table of Contents

Project Overview ***3***

Technology Stack ***3***

Problem Statement and Solution Approach ***4***

Implementation Highlights ***5***

Evaluation ***6***

Conclusion ***7***

References ***7***

Project Overview

This project uses Python and natural language processing (NLP) techniques to perform sentiment analysis on Amazon product reviews. By leveraging the power of the NLTK library and the Transformers framework, the goal is to develop a model capable of accurately classifying reviews as positive or negative. This project showcases the practical application of NLP and machine learning in analysing real-world text data, which has immense value in various domains such as e-commerce, customer feedback analysis, and brand monitoring.

The project repository can be found at: https://github.com/carlos-alves-one/Amazon-Review-NLP

The repository contains the necessary code, datasets, and documentation to reproduce the sentiment analysis pipeline. It is a comprehensive resource for understanding the project's implementation details and facilitates collaboration and knowledge sharing within the NLP community.

Technology Stack

- Python
- NLTK (Natural Language Toolkit)
- Transformers framework
- Git (version control system)

This project aims to demonstrate an understanding of NLP concepts, proficiency in Python programming, and the ability to apply machine learning models to real-world text data. It will utilize advanced NLP libraries and frameworks like NLTK and Transformers and gain hands-on experience in pre-processing text data, feature extraction, model training, and evaluation. In this project, Git was crucial for maintaining a well-structured and organized development workflow. It facilitated code review processes and provided a reliable mechanism for tracking and managing different project versions.

Problem Statement and Solution Approach

This project aims to analyse Amazon customer reviews and classify them into binary sentiment categories: positive or negative. This involves processing natural language data to capture the sentiment expressed by customers regarding their experiences with Amazon products. The challenge lies in accurately identifying and categorising the sentiment of each review, considering the nuances of human language, such as sarcasm, context, and the presence of negations.

Solution Approach:

The approach to this problem is structured into several key steps:

1. Data Collection: Acquire the dataset of Amazon reviews, ensuring it includes review texts and associated ratings.

2. Data Pre-processing:

- Select the relevant columns from the dataset necessary for sentiment analysis (i.e., the review text and ratings).
- Handle any missing values to maintain data integrity.
- Map review ratings to a binary sentiment label, where ratings above three are considered positive (1) and those below as negative (0).

To simplify the data for analysis, Clean the review text by tokenising, converting to lowercase, removing punctuation, and excluding stop words.

3. Text Normalization:

- Apply lemmatisation and stemming to reduce words to their base or root form, enhancing the uniformity of the dataset.
- Implement negation handling to ensure that phrases like "not good" are accurately represented and do not skew sentiment analysis.

4. Text Vectorization:

- Use TF-IDF vectorisation to transform the pre-processed text into numerical vectors that highlight the importance of each term in the context of the document corpus.

- Explore advanced embeddings like Word2Vec and BERT to capture semantic relationships between words and contextual nuances.

5. Advanced Pre-processing and Feature Engineering:

- Remove custom stop words that are not informative for sentiment analysis.
- Handle synonyms to consolidate similar meanings and reduce feature space.
- Incorporate n-grams to capture multi-word expressions that might be crucial for sentiment analysis.
- Apply feature scaling to standardise the range of independent variables, improving model performance.

6. Model Building and Evaluation:

- Train machine learning models using vectorised text data with Logistic Regression.
- Evaluate model performance using accuracy and F1 score to assess predictive capabilities and the confusion matrix to understand classification errors.
- Utilise ROC curves and AUC metrics to quantify the models' ability to differentiate between sentiment classes.

By employing these NLP methods and machine learning techniques, the project aims to create a robust sentiment analysis model to help businesses understand customer feedback and enhance customer satisfaction.

Implementation Highlights

The sentiment analysis project for Amazon customer reviews involved crucial steps such as data pre-processing, text normalization, text vectorization, model building and evaluation, advanced pre-processing and feature engineering. The project aims to accurately classify customer reviews into positive and negative sentiment categories, providing valuable insights for businesses to understand customer feedback and improve their products and services.

However, the project faced several unique challenges. Sarcasm and contextual nuances in customer reviews can be difficult to detect and interpret correctly. Negative phrases like "not good" can invert a statement's sentiment, requiring careful handling to ensure accurate

sentiment classification. Additionally, Amazon reviews may contain domain-specific terminology, abbreviations, or slang that can be challenging for generic NLP models to understand.

Advanced NLP techniques were employed to overcome these challenges. BERT embeddings can capture the contextual meaning of words and phrases, enabling a better understanding of sarcasm and underlying sentiment. Implementing negation handling techniques, such as reversing the sentiment polarity of words following negation terms, can help mitigate the impact of negation and improve sentiment classification accuracy.

Another significant challenge is the potential imbalance between the dataset's positive and negative sentiment classes, which can lead to biased model predictions. Techniques such as oversampling the minority class or under sampling the majority class can help mitigate the impact of class imbalance and improve the model's ability to classify both sentiment categories accurately, which can be done in future research.

By addressing these challenges through careful data pre-processing, advanced NLP techniques, and model optimization, the sentiment analysis project can effectively classify Amazon customer reviews into positive and negative sentiment categories. This enables businesses to gain valuable insights from customer feedback, identify areas for improvement, and enhance overall customer satisfaction.

Evaluation

The sentiment analysis project on Amazon reviews demonstrates the effectiveness of various natural language processing techniques and machine learning models in accurately classifying sentiment from text data. The project's systematic approach, starting with data pre-processing, text normalization, and vectorization, followed by model building and evaluation, highlights the importance of each step in achieving optimal results.

Among the models evaluated, the BERT embedding approach achieved the highest performance, with an accuracy of 92.61% and an F1-score of 92.02%. This showcases the

power of advanced pre-trained language models in capturing rich semantic information and context, leading to superior sentiment classification accuracy.

The TF-IDF and Word2Vec vectorization approaches, combined with logistic regression, also yielded promising results, with accuracies of 88.22% and F1-scores of 83.97%. However, the confusion matrices revealed that these models needed help correctly classifying negative sentiment reviews compared to the BERT approach.

The project also explored advanced text pre-processing and feature engineering techniques, such as custom stop word removal, synonym handling, n-grams, and feature scaling. These techniques improved the logistic regression model's performance, as demonstrated by the ROC curve analysis and an AUC value of 92%.

Conclusion

In conclusion, this sentiment analysis project on Amazon reviews showcases the effectiveness of various NLP techniques and machine learning models in accurately classifying sentiment. The BERT embedding approach emerges as the top-performing model, while advanced pre-processing and feature engineering techniques significantly enhance the logistic regression model's performance. The project emphasises the importance of selecting appropriate methods and evaluation metrics to achieve optimal results in sentiment classification tasks. It provides valuable insights into applying NLP and machine learning in real-world scenarios, such as analysing customer feedback and improving product development strategies. This project demonstrates a comprehensive application of NLP techniques for sentiment analysis. It opens avenues for significant improvements and expansions, promising enhanced understanding and utilisation of customer feedback in business strategies.

References

- Natural Language Processing: Dr. Akshi Kumar, Senior Lecture in MSc. in Data Science and AI at Goldsmiths University of London
- Natural Language Processing with Python - Steven Bird, Ewan Klein, Edward Loper
- Python Text Processing with NLTK 2.0 Cookbook - Jacob Perkins, 2010