Activity based

Project Report

Natural Language Processing

Project I

Submitted to Vishwakarma University, Pune

By

Abdul Basit Mulla

SRN No : 202100792

Div : E

Fourth Year Engineering

Faculty Incharge:- Prof. Pavitha Noji

Date Of Project : 03-10-2024

Department of Computer Engineering

Faculty of Science and Technology

Academic Year

2024-2025 Term-I

NLP : Project I

Project Name : Sentiment Analysis on Book Reviews

Abstract

This report presents an analysis of sentiments expressed in Amazon Kindle reviews using a dataset comprising of 10k+ user-generated reviews. The primary objective was to classify reviews into positive and negative sentiments based on the ratings provided. Various machine learning models were employed, including Logistic Regression, Support Vector Machine (SVM), and Convolutional Neural Networks (CNN). This project involved data cleaning, exploratory data analysis (EDA), feature extraction, and model training and evaluation to optimize performance and ensure robust sentiment classification.

Introduction

Sentiment analysis is a vital component of natural language processing (NLP) that aims to determine the emotional tone behind a series of words, helping businesses and researchers understand public opinion. This project focuses on analyzing Amazon Kindle reviews, leveraging machine learning techniques to classify sentiments into positive and negative categories. The goal is to provide a detailed understanding of customer sentiments and identify patterns in reviews, thereby aiding authors and publishers in improving their offerings.

Implementation:

Dataset Exploration

- **Dataset Source**: Amazon Kindle Reviews (Kaggle)
- **Total Reviews**: 12,000
- **Sentiment Distribution**:
- **Note** : Initially, we included reviews rated as 3 stars as positive. However, to achieve better binary classification and handle class imbalance, we later ignored the neutral ratings (3 stars). This adjustment resulted in 8,000 positive reviews and 4,000 negative reviews being reduced to 6,000 positive and 4,000 negative reviews, with 2,000 neutral ratings removed.
    - Positive Sentiments: 6,000 reviews
    - Negative Sentiments: 4,000 reviews

- **Features**:
  - reviewText: The actual review content.
  - summary: A brief summary of the review.
  - rating: User's rating from 1 to 5.

Exploratory Data Analysis (EDA) and Data Cleaning

1. **Missing Values**: Checked for missing values across all columns. No significant missing data was found, ensuring a complete dataset for analysis.
2. **Duplicate Reviews**: Identified and removed any duplicate entries to ensure data integrity.
3. **Class Imbalance**: Acknowledged the potential for class imbalance but decided to proceed with the analysis, given the significant difference in positive and negative review counts.
4. **Sentiment Distribution Visualization**:
   a. **Bar Plot**: Visualized the distribution of sentiments to highlight the imbalance.
   b. **Word Clouds**: Generated separate word clouds for positive and negative sentiments to identify frequently used words in each category.
5. **Textual Analysis**:
   - The most common words in positive reviews often included "love," "great," and "best," indicating satisfaction, while negative reviews featured words such as "disappointing," "bad," and "waste," suggesting dissatisfaction.

Text Preprocessing

To prepare the text data for modeling, the following preprocessing steps were undertaken:

1. **Cleaning**:
   - Removed special characters, numbers, and punctuation.
   - Converted all text to lowercase to maintain uniformity.
2. **Tokenization**:
   - Split the reviews into individual words for further analysis.
3. **Stopwords Removal**:
   - Removed common stopwords (e.g., "the," "is," "and") to focus on meaningful words.
4. **Lemmatization**:
   - Implemented lemmatization to reduce words to their base or root form (e.g., "running" to "run").
5. **Vectorization**:
   - Employed TF-IDF (Term Frequency-Inverse Document Frequency) with n-grams to create a matrix representation of the text data, capturing both the frequency and importance of words.

Model Training

The dataset was split into training and testing sets (80% training, 20% testing) for the machine learning models. Additionally, a separate 70/30 split was utilized for the CNN model to maximize training data.

1. **Logistic Regression**:
    - **Hyperparameters**: Optimized using GridSearchCV, achieving the best parameters with C = 10, Penalty = 'l2', Solver = 'liblinear'.
- **Accuracy**: 88%
- **Classification Report**:
    - **Negative (0)**:
        - Precision: 0.87
        - Recall: 0.83
        - F1-Score: 0.85
    - **Positive (1)**:
        - Precision: 0.89
        - Recall: 0.92
        - F1-Score: 0.91
    - Confusion Matrix: Showed balanced predictions with a few false negatives.

Logistic Regression achieved an accuracy of 88%, matching the performance of the SVM model. Its precision for positive reviews was slightly higher at 0.89, with a recall of 0.92, indicating strong performance in identifying positive sentiments. However, its precision and recall for negative reviews (0.87 and 0.83, respectively) were slightly lower compared to SVM.

2. **Support Vector Machine (SVM)**:
    - **Hyperparameters**: Tuned for optimal performance.
    - **Accuracy**: 88%
    - **Classification Report**:
    - **Negative (0)**:
        - Precision: 0.87
        - Recall: 0.82
        - F1-Score: 0.85
    - **Positive (1)**:
        - Precision: 0.88
        - Recall: 0.92
        - F1-Score: 0.90
    - Confusion Matrix: Illustrated minor improvements in true positive rates compared to Logistic Regression.

The SVM model produced an accuracy of 88%, making it the top-performing model in terms of overall accuracy. It showed balanced performance with a high precision of 0.88 and recall of 0.92 for positive reviews, indicating that it was effective in

identifying positive sentiments. The model also maintained reasonable performance for negative reviews, achieving a precision of 0.87 and a recall of 0.82.

3. **Convolutional Neural Network (CNN)**:
    - **Architecture**: Designed with embedding, convolutional layers, batch normalization, max pooling, and dropout layers to enhance model robustness and prevent overfitting.
    - **Results**:
    - **Test Accuracy**: 86.07%
    - **Classification Report**:
    - **Negative**:
        - Precision: 0.84
        - Recall: 0.80
        - F1-Score: 0.82
    - **Positive**:
        - Precision: 0.87
        - Recall: 0.90
        - F1-Score: 0.89
    - Confusion Matrix: Showed slight improvements in identifying positive sentiments compared to traditional models.

The CNN model achieved a decent accuracy of 86.07%. It performed better at identifying positive reviews with a precision of 0.87 and recall of 0.90, indicating a strong ability to correctly classify positive sentiments. However, it had a lower recall for negative reviews (0.80), suggesting some false negatives where negative sentiments were misclassified as positive.

Summary of Model Comparisons

- **Accuracy**: Both SVM and Logistic Regression achieved the highest accuracy of 88%, while CNN followed closely with 86.07%.
- **Precision**: CNN had lower precision for negative sentiments compared to both SVM and Logistic Regression, which performed similarly in this regard.
- **Recall**: SVM exhibited the best balance in recall for both sentiment classes, effectively capturing a higher proportion of true positive reviews while maintaining a reasonable level of false negatives.
- **F1-Score**: The SVM model also led in F1-scores for both classes, reflecting its ability to achieve a balance between precision and recall.

Overall, while all three models performed well, the SVM model emerged as the most effective for our sentiment analysis, demonstrating the highest accuracy, precision, and recall metrics. This suggests that the SVM approach is better suited for distinguishing between positive and negative sentiments in the dataset.

Conclusion

The sentiment analysis project effectively classified Amazon Kindle reviews into positive and negative sentiments using various machine learning models. The CNN outperformed the traditional models, affirming the efficacy of deep learning approaches in sentiment analysis tasks. The findings suggest that sentiment analysis can provide valuable insights into customer opinions, which can aid businesses in understanding user satisfaction and improving their products. Future work may involve enhancing the CNN architecture, exploring more advanced techniques such as BERT or LSTM, and utilizing additional features to further improve model performance.