WIDEBOT

Sentiment Analysis using NLP

OVERVIEW

The project focuses on sentiment analysis of Amazon Kindle book reviews to classify reviews into three categories: negative, neutral, and positive. The aim is to develop a model that accurately reflects customer sentiment based on their reviews.

1. Data Description

Dataset Source: The dataset is sourced from Amazon Kindle Book Reviews and is provided as a CSV

Features:

- **Review Text: The textual content of the Kindle book reviews.**
- **Rating: Numerical sentiment score assigned to each review, ranging from 0 to 5.**

Data Preprocessing:

- **Sentiment ratings are mapped to categorical labels: 0 = negative, 1 to 3 = neutral, 4 to 5 = positive.**
- **The dataset is divided into training and testing sets to evaluate the model's performance. The split ratio is 60% training and 40% testing.**

2. Baseline Experiments

Goal :
The goal of the baseline experiments is to establish a reference performance using a simple text classification approach.

Initial Experiments:

- **Model Used: Multinomial Naive Bayes.**
- **Feature Extraction: TF-IDF (Term Frequency-Inverse Document Frequency) to convert text data into numerical features.**
- **Evaluation Metrics: Accuracy, Precision, Recall, and F1-Score.**

Results:

- **Accuracy: 0.82**
- **Neutral Sentiments: The model achieved a precision of 0.82 and a recall of 0.83, resulting in an F1-score of 0.83. This indicates a good balance between identifying neutral sentiments and minimizing false positives.**
- **Positive Sentiments: The precision was 0.83, and recall was 0.81, with an F1-score of 0.82, showing robust performance in identifying positive sentiments.**

Conclusion: The initial experiments provide a baseline performance for the sentiment analysis task. The Multinomial Naive Bayes classifier with TF-IDF features gives a good starting.

3 Overall Conclusion

Summary:

The sentiment analysis project successfully classified Kindle book reviews into negative, neutral, and positive categories. The Multinomial Naive Bayes model with TF-IDF feature extraction provided a solid baseline

Table 1: Dataset Description

Feature	Description
Source	Amazon Kindle Book Reviews
Number of Instances	4,800 reviews
Features	- Review Text: The content of the review
	- Rating: Numerical score from 0 to 5
Sentiment Labels	Negative (0), Neutral (1-3), Positive (4-5)
Data Split	Training: 80%, Testing: 20%

Class	Precision	Recall	F1-Score	Support
Neutral	0.82	0.83	0.83	2413
Positive	0.83	0.81	0.82	2387
Accuracy			0.82	4800
Macro Avg	0.82	0.82	0.82	4800
Weighted Avg	0.82	0.82	0.82	4800

Tools and Libraries Used

1. **Python**: The primary programming language used for implementing the sentiment analysis pipeline.
2. **Pandas**: Library used for data manipulation and preprocessing, including loading and transforming the dataset.
3. **Scikit-Learn**: Machine learning library used for:
    - **Feature Extraction**: `TfidfVectorizer` for converting text data into numerical features.
    - **Model Training**: `MultinomialNB` (Multinomial Naive Bayes) for classifying sentiment.
    - **Evaluation**: Metrics such as `accuracy_score` and `classification_report` for assessing model performance.
    - **Data Splitting**: `train_test_split` for dividing the dataset into training and testing sets.

External Resources or Pre-trained Models Used

- **Amazon Kindle Book Reviews Dataset**: The dataset contains user reviews and ratings for Kindle books. [Amazon Kindle Book Review for Sentiment Analysis (kaggle.com).]

Reflection Questions:

1: It's hard to understand all of the techniques because it’s new for me but I had the chance to cover it all.

2: the importance of data handling and how small things can be so affected on the hole process