Sentiment Analysis Using Multiple Approaches

Rahul Mistry, Sidhant Rohatgi, Sithu Soe, Aayush Motiani

1. Introduction

Sentiment analysis is a critical area of natural language processing (NLP) that aims to determine the sentiment or emotional tone behind a piece of text. It has applications in various domains, including product reviews, social media analysis, customer feedback systems, and political sentiment tracking. The objective of this project is to classify customer reviews into three sentiment categories: **Positive**, **Neutral**, and **Negative**, based on textual data.

Customer reviews often serve as a rich source of information for businesses, offering direct insights into user satisfaction and product perception. By understanding the polarity of these reviews, organizations can refine their strategies to address customer concerns and enhance their offerings.

The primary challenge in sentiment analysis lies in the complexity and variability of human language. Reviews often contain nuanced expressions, sarcasm, or mixed sentiments, making classification a non-trivial task. Additionally, imbalanced datasets—where one sentiment dominates—add another layer of complexity to the task.

This project explores multiple approaches to sentiment analysis using machine learning. Specifically, we investigate:

1. **TF-IDF with N-grams**: A frequency-based method that captures the importance of words and phrases.
2. **Word2Vec embeddings**: A semantic representation technique that encodes contextual relationships between words.
3. **Combined Features**: A hybrid approach integrating TF-IDF and Word2Vec features for improved performance.

Through these approaches, we aim to evaluate and compare the effectiveness of different feature extraction methods for sentiment classification.

2. Dataset and Exploratory Analysis

Dataset Description

The dataset consists of 70,000 customer reviews from an e-commerce platform. Each review is associated with:

- **Text**: A string of text describing the customer's opinion about a product or service.
- **Rating**: A numerical score between 1 and 5, where higher scores indicate positive experiences.
- **Sentiment**: Derived from the rating attribute:
    - **Positive** (ratings 4–5),
    - **Neutral** (rating 3),
    - **Negative** (ratings 1–2).

The dataset includes diverse reviews, ranging from single-word comments like "Excellent!" to multi-sentence descriptions discussing product pros and cons. This diversity presents opportunities and challenges for the sentiment classification task.

Exploratory Data Analysis

Class Distribution

An initial analysis reveals a significant class imbalance:

- **Positive**: 70% (49,000 reviews),
- **Neutral**: 10% (7,000 reviews),
- **Negative**: 20% (14,000 reviews).

This imbalance poses a challenge for the classification model, particularly in accurately predicting Neutral sentiments.

Text Length Analysis

The average length of reviews is approximately 100 words. Negative reviews tend to be shorter and more direct, often expressing dissatisfaction concisely. In contrast, Neutral reviews are typically longer and nuanced, reflecting ambivalence or mixed opinions. Positive reviews vary significantly in length, ranging from brief exclamations to detailed praise.

Frequent Words and Phrases

A unigram analysis shows that Positive reviews frequently include words like "good," "love," and "excellent," while Negative reviews commonly feature terms like "bad," "poor," and "disappointed." Bigram and trigram analyses reveal phrases like "highly recommend" (Positive) and "not worth it" (Negative).

These findings motivated the use of n-grams and semantic embeddings to capture both the frequency and context of words in reviews.

3. Predictive Task and Features

Task Definition

The goal is to classify reviews into three sentiment categories. The predictive task involves:

1. Extracting meaningful features from textual data.
2. Training machine learning models to predict sentiment labels.
3. Evaluating the performance of different approaches.

Feature Engineering

1. **TF-IDF with N-grams**:
   - Captures term importance using unigrams, bigrams, and trigrams.
   - Highlights frequently used words and phrases indicative of sentiment.
2. **Word2Vec Embeddings**:
   - Encodes semantic relationships between words.
   - Represents each review as the average of its word embeddings.
3. **Combined Features**:
   - Integrates the strengths of both approaches.
   - Balances frequency-based signals (TF-IDF) with contextual understanding (Word2Vec).
   -

Evaluation Metrics

The models are evaluated using:

- **Accuracy**: Percentage of correctly predicted reviews.
- **Precision, Recall, F1-Score**: To account for the class imbalance.
- **Confusion Matrix**: To visualize misclassifications.

4. Methodology

Model Selection

Logistic Regression was chosen for its simplicity, efficiency, and interpretability. It performs well on text classification tasks, particularly when combined with robust feature extraction methods.

Approaches

1. TF-IDF with Logistic Regression

TF-IDF (Term Frequency-Inverse Document Frequency) assigns weights to terms based on their frequency in a document and their rarity across the corpus. Using n-grams (up to trigrams) allows the model to capture contextual phrases, improving sentiment prediction.

2. Word2Vec with Logistic Regression

Word2Vec generates dense vector representations of words, capturing semantic relationships. By averaging the embeddings of all words in a review, we create document-level representations for classification.

3. Combined Features

The combined approach merges scaled TF-IDF features and Word2Vec embeddings. Weighted contributions (70% TF-IDF, 30% Word2Vec) were used to balance the complementary strengths of these methods.

Optimization

1. **Hyperparameter Tuning**:
   - Regularization strength (C) and solver parameters for Logistic Regression were optimized using GridSearchCV.
2. **Feature Scaling**:
   - Word2Vec embeddings were standardized to match TF-IDF's scale.

5. Results and Discussion

Performance Metrics

Model	Acc	Precision (Neutral)	Recall (Neutral)	F1-Score (Neutral)
TF-IDF	84%	0.26	0.04	0.07
Word2Vec	81%	0.22	0.03	0.05
Combined	86%	0.28	0.12	0.18

Key Findings

1. **TF-IDF**:
   - Performed well on Positive and Negative sentiments.
   - Struggled with Neutral due to vocabulary overlap with Positive reviews.
2. **Word2Vec**:
   - Captured semantic nuances but was less effective on short reviews.
3. **Combined Features**:
   - Achieved the highest accuracy by leveraging both frequency-based and semantic information.
   - Improved Neutral sentiment classification, though challenges remain.

Visualization

- Confusion matrices illustrate improved predictions for Neutral sentiments in the combined model.
- Feature importance analysis highlights the contribution of n-grams and key semantic vectors.

6. Related Literature

Aayush and Sidhant please clutch

7. Conclusion

This project successfully classified customer reviews into three sentiment categories using TF-IDF, Word2Vec, and combined features. The combined model achieved the highest accuracy (86%) by integrating frequency-based and semantic signals. Future work will explore deep learning models like BERT to address the challenges of imbalanced data and nuanced sentiments.