Sentiment Analysis of Movie Reviews

Divya Dhole
University of Arizona
Divyadhole@arizona.edu

Abstract

This report presents a sentiment analysis of movie reviews using the IMDb movie reviews dataset from the NLTK corpus. The primary goal is to classify reviews into positive and negative sentiments using machine learning techniques, specifically Logistic Regression and Support Vector Machine (SVM).

1 Introduction

This report presents a sentiment analysis of movie reviews using the IMDb movie reviews dataset from the NLTK corpus. The primary goal is to classify reviews into positive and negative sentiments using machine learning techniques, specifically Logistic Regression and Support Vector Machine (SVM).

2 Dataset

The dataset used for this analysis is the **IMDb movie reviews dataset**, which consists of 2,000 movie reviews (1,000 positive and 1,000 negative). Each review is labeled according to its sentiment, either 'pos' for positive or 'neg' for negative.

3 Methodology

3.1 Data Preparation

- Loaded the IMDb movie reviews dataset from NLTK. - Preprocessed the reviews by converting the text to lowercase, removing stopwords, and using TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction.

3.2 Model Training

- Split the dataset into training (80%) and testing (20%) sets. - Trained two different machine learning models:

- **Logistic Regression**
- **Support Vector Machine (SVM)**

3.3 Evaluation Metrics

- Evaluated model performance using accuracy, precision, recall, and F1-score.

4 Implementation

4.1 Data Preprocessing

The following steps were performed to preprocess the data:

- **Stopword Removal:** Common English words (stopwords) were removed to reduce noise in the data. - **TF-IDF Vectorization:** Transformed the reviews into TF-IDF features to represent the importance of words in the documents.

4.2 Model Training and Evaluation

4.2.1 Logistic Regression Model

```
# Train Logistic Regression model
logreg_classifier = LogisticRegression(max_i
logreg_classifier.fit(X_train_tfidf, y_train

# Predict on the test data
y_pred_logreg = logreg_classifier.predict(X_

# Evaluate Logistic Regression model perform
accuracy_logreg = accuracy_score(y_test, y_p
```

- **Accuracy:** 83.75% - **Classification Report:**

	precision	recall	f1-score
neg	0.84	0.83	0.84
pos	0.84	0.84	0.84
accuracy			0.84
macro avg	0.84	0.84	0.84
weighted avg	0.84	0.84	0.84

4.2.2 Support Vector Machine (SVM) Model

```
# Train SVM model
svm_classifier = SVC(kernel='linear')
svm_classifier.fit(X_train_tfidf, y_train)
```

```
# Predict on the test data
y_pred_svm = svm_classifier.predict(X_test_tfidf)

# Evaluate SVM model performance
accuracy_svm = accuracy_score(y_test, y_pred_svm)
```

- **Accuracy:** 82.25% - **Classification Report:**

```
              precision    recall  f1-score   support

         neg       0.82      0.82      0.82       199
         pos       0.82      0.83      0.82       201

    accuracy                           0.82       400
   macro avg       0.82      0.82      0.82       400
weighted avg       0.82      0.82      0.82       400
```

5 Results and Discussion

The results indicate that both models performed well on the sentiment classification task. The Logistic Regression model achieved a higher accuracy (83.75%) compared to the SVM model (82.25%). Both models demonstrated similar performance in terms of precision, recall, and F1-score, highlighting the reliability of the Logistic Regression model in classifying movie reviews.

5.1 Key Observations

- **Logistic Regression** had a balanced precision and recall for both classes, indicating its effectiveness in distinguishing between positive and negative sentiments. - **SVM** also performed well but had slightly lower accuracy compared to Logistic Regression, suggesting that while SVM is a powerful algorithm, it may require more tuning for optimal performance in this specific task.

6 Conclusion

This analysis successfully classified movie reviews into positive and negative sentiments using Logistic Regression and SVM models. While both models yielded promising results, Logistic Regression proved to be the more effective model in this case. Future work may involve exploring additional preprocessing techniques, hyperparameter tuning, and experimenting with other advanced models to further improve sentiment classification performance.