Sentiment Analysis Report

Rajiv Chaudhary

1. Abstract

Sentiment analysis, also known as opinion mining, is a crucial task in natural language processing (NLP) that involves determining the sentiment expressed in textual data. This project aims to classify text into positive, negative sentiments using various machine learning models and embedding techniques. The dataset underwent extensive preprocessing.The preprocessed text data was then converted into numerical representations using three different feature extraction methods: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec embeddings.

Several models were trained and evaluated, including traditional classifiers like Naive Bayes, Random Forest, and gradient boosting models (XGBoost, CatBoost, LightGBM), as well as deep learning approaches such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks. The models' performances were assessed using accuracy scores, and results indicate that different models excel under different vectorization techniques. CatBoost achieved the highest accuracy across multiple vectorization methods, while LSTM performed best with Word2Vec embeddings.

2. Introduction

Sentiment analysis, also known as opinion mining, is the process of computationally identifying and categorizing opinions expressed in text to determine whether the sentiment is positive, negative, or neutral.

Traditional sentiment analysis techniques relied on rule-based systems and lexicon-based methods, which required extensive manual effort and predefined sentiment dictionaries. However, with advancements in machine learning and deep learning, sentiment analysis has evolved into a more automated and efficient process. Machine learning algorithms can generalize patterns in data and classify text with higher accuracy. Moreover, deep learning models such as Recurrent Neural Networks (RNNs) and Transformers have further improved sentiment analysis by capturing contextual relationships and semantic meanings in text.

This study explores various sentiment analysis models, ranging from traditional algorithms like Naive Bayes and Random Forest to advanced deep learning architectures like LSTM.

3. Dataset

3.1. IMDB Movie Reviews dataset

we used the IMDB Movie Reviews dataset, which consists of 50,000 movie reviews labeled as either positive or negative. This dataset is widely used in sentiment analysis research due to its balanced class distribution and real-world relevance. Each review contains free-text movie critiques, making it a challenging yet insightful dataset for sentiment classification tasks.

The dataset was split into 80% training data and 20% testing data to ensure robust evaluation of the models. The processed text was then transformed into numerical representations using BoW, TF-IDF, and Word2Vec embeddings.

3.2. Data Preprocessing

The text data underwent preprocessing to remove noise and improve the quality of input features. We applied various transformations.

such as lowercasing, removal of URLs and special characters, tokenization, and lemmatization.

- **lowercasing**
- **removal of URLs and tags**
- **removing special characters**
- **removing stop words**
- **tokenization**
- **lemmatization**

4. Text Vectorization

The text data was converted into numerical representations using the following methods:

- Bag of Words (BoW)
- Term Frequency-Inverse Document Frequency (TF-IDF)
- Word2Vec

5. Model Selection

The following models were implemented:

- **Multinomial Naive Bayes (MultinomialNB)**
- **Gaussian Naive Bayes (GaussianNB)**
- **Random Forest**

- **XGBoost**
- **CatBoost**
- **LightGBM**
- **Artificial Neural Network (ANN)**
- **Long Short-Term Memory (LSTM)**

6. Results

The models were evaluated using different embedding techniques. The results are shown in Table 1.

Model	BoW	TF-IDF	Word2Vec
MultinomialNB	0.8437	0.8531	nan
GaussianNB	0.7137	0.777	0.7544
Random Forest	0.8432	0.8428	0.8255
XGBoost	0.8511	0.8541	0.8436
CatBoost	0.871	0.8662	0.8574
LightGBM	0.8561	0.8566	0.8455
ANN	0.8498	0.8648	0.8509
LSTM	nan	nan	0.8671

Table 1. Performance Metrics of Different Models with Various Embedding Techniques

7. Conclusion

This study explored various machine learning models and text representation techniques for sentiment analysis. The CatBoost model showed the highest performance across multiple vectorization methods. The LSTM model performed best with Word2Vec embeddings.