Sentiment Analysis on IMDB Dataset

Zhe Cao[1,a]

[1]*Institute of Computer Science, Nanjing University, 163 Xianlin Road, Nanjing, China*
a. 231220100@smail.nju.edu.cn

1. Introduction

Sentiment analysis is a key task in Natural Language Processing, where the goal is to determine the sentiment polarity of a text. The IMDB movie review dataset is widely used for sentiment analysis, and it provides a collection of movie reviews labeled with their respective sentiment. In this paper, we explore how to perform sentiment analysis on the IMDB dataset. In this paper, we use the IMDB movie review dataset, which contains a set of 25k highly polar movie reviews for training and 25k for testing. The dataset is available at *https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews*.

2. Literature Review

Sentiment analysis has become an essential task in Natural Language Processing (NLP) and has significant applications across various domains, such as social media analysis, product reviews, movie critiques, and customer feedback.

2.1. Application of NLP in Analyzing Sentiment

Social media platforms like Weibo, have emerged as rich sources of public opinion. NLP techniques have been extensively applied to analyze sentiment in social media posts. For example, sentiment analysis of tweets has been used to assess public opinion on political events, track consumer sentiment about products, and even analyze social movements.

2.1.1. TF-IDF

This technique assigns a weight to each word in a document based on its frequency and its importance relative to the entire corpus. We use the technique to capture key features of the text.

2.2. Machine Learning Method

Machine learning techniques have revolutionized natural language processing by enabling more accurate models. These methods enhance understanding, improve accuracy, and enable real-time language processing.

2.2.1. Naive Bayes

A probabilistic classifier that assumes feature independence. Despite its simplicity, Naive Bayes has been widely used for sentiment analysis tasks due to its speed and performance on smaller datasets.

2.2.2. Support Vector Machine

SVM is a powerful classification technique that is particularly effective in high-dimensional spaces. It works by finding the hyperplane that best separates classes in a feature space, making it a popular choice for sentiment classification tasks.

3. Methodology and Analysis

In study, the IMDB movie review dataset is used to perform sentiment analysis. The following steps were undertaken:

3.1. Text Preprocessing

At first, we need to load the data from the dataset. And the next, data from the IMDB dataset was preprocessed using the steps contains lowercasing, tokenization, stopword removal, punctuation removal and stemming.

3.2. Modeling

3.2.1 Model Train

After preprocessing, the data was vectorized using the TF-IDF technique to transform the textual data into numerical features. These features were then fed into naïve bayes model and a SVM model, which was chosen for its simplicity and effectiveness in binary classification tasks with large-scale text.

3.2.2 Model Evaluation

After training the models, we evaluate their performance using following metrics. The model's performance was evaluated using **accuracy**, which measures the proportion of correctly classified reviews. The dataset was split into training and test sets.

3.3. Outcome

By using cross validation method, we get the models classification acc. The naïve bayes model could access 85.5% and SVM model could get 88% acc by contrast. They both have a great performance in the sentiment analysis task.

4. Research Proposal

To improve the current analysis, we could integrate more high-level NLP techniques to make the analysis better, things like using deep learning structure like BERT(Bidirectional Encoder Representations from transformer) and so on. And the training time is a critical problem, we trained for 2 hours with one AMD CPU upon the 50k dataset, so the following we could find a way to train using GPU to accelerate compute.

References

[1] *M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428..*

[2] *Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation.* ***28*** *(1): 11–21.*