Movie Reviews: A machine learning project

D. Ligari[1]

[1] *University of Pavia, Department of Computer Engineering (Data Science), Pavia, Italy*

Date: May 1, 2023
Contact: davide.ligari01@universitadipavia.it

Abstract— This report provides a comprehensive analysis of the steps taken to develop a machine learning model for classifying film reviews. The dataset used in this study is Andrew Maas's *Large Movie Review Dataset*, which includes 50,000 labelled film reviews as either positive or negative. Prior to training the model, a preprocessing phase was conducted, which involved creating a vocabulary consisting of all words in the dataset, and a bag-of-words (BOW) approach where the number of occurrences of each word in each review was counted. The BOW approach was subsequently used to train the model.
Different models were then created and compared, incorporating various preprocessing techniques, including stemming and the removal of common and meaningless words from the vocabulary.
The report also includes the use of different classifiers, such as logistic regression and multinomial Naive Bayes.

Keywords— Machine Learning • Classification • Sentiment Analysis • Logistic Regression • Multinomial Naive Bayes

1. Introduction

In recent years, the exponential growth of digital media and the internet has led to an unprecedented amount of content being produced and consumed on a daily basis. One of the most popular forms of digital media is film, with millions of reviews available online from a variety of sources. With this vast amount of data available, it has become increasingly challenging to manually process and classify this information.
Machine learning provides a powerful tool for automating the classification of text data, including film reviews. By using large datasets and advanced algorithms, machine learning models can accurately classify reviews based on their content, sentiment, and other criteria.
In this report, we present a thorough analysis of the steps taken to create a machine learning model for classifying film reviews.

2. Dataset

The dataset used in this study is Andrew Maas's *Large Movie Review Dataset*, which includes 50,000 labelled film reviews equally distributed in the two classes - positive and negative. The dataset is already divided into a training set of 25,000 reviews, a validation set of 12,500, and a test set of 12,500, which allows for the evaluation of the model's performance on unseen data. The data set has been introduced in the paper: "Learning Word Vectors for Sentiment Analysis," by Andrew L. Maas et al., In addition, a subset of the training set is also available for faster experimentation. The subset contains a smaller number of reviews than the full training set, which can help reduce the computational resources required for training the model.

3. Model creation

The initial step in building the model is to construct the vocabulary, which is a comprehensive list of all words present in the dataset. The vocabulary was generated by processing all training reviews, removing any punctuation, and adding each unique word to the vocabulary only once.
Afterward, the vocabulary was sorted alphabetically and filtered to retain only a subset of the most relevant words. The size of the vocabulary is a crucial parameter that can be adjusted to enhance the model's performance.
Once the vocabulary has been created, the next step is to extract meaningful features from the dataset. In this study, a bag-of-words (BOW) approach has been used, which counts the number of occurrences of each word in each review. Thorough the bag of words, the reviews are represented in a structured and quantitative way, making it possible to train machine learning models to accurately classify them.
Now the data are ready to be used to train the model. In this study,the multinomial Naive Bayes classifier has been used. Given a BoW feature vector x the multinomial NB model predicts the class $\hat{y}$ as follows:

$$\hat{y} = \underset{y\in\{0,1\}}{argmax} \sum_{j=0}^{n-1} x_j \log \pi_{y,j} + \log P(y) \tag{1}$$

where $\pi_{y,j}$ represents the probability that a randomly selected word from a document belonging to class y is the j-th word in the vocabulary. The term P(y) refers to the prior probability for class y.

a. Variants

In this study, we analyzed different variants of our model by training it with various vocabularies. In the first variant, we removed the most common, meaningless words such as articles, prepositions, and conjunctions, among others. In the second one, the stemming technique has been applied, to reduce words to their root form, in order to reduce the vocabulary size, and so, improve the performance of the model. In the third variant, both techniques has been combined, to create a more comprehensive vocabulary.

4. Model evaluation

a. Effect of the vocabulary size

The vocabulary size is a crucial parameter that affects the performance of the model. A vocabulary that is too small may not contain enough information to accurately classify reviews, while a vocabulary that is too large may include irrelevant words that negatively impact the model's performance. When the vocabulary size is too large, the model may assign a high weight to meaningless words, which can reduce its ability to identify the most relevant ones.
However, as shown in Figure 1, increasing the vocabulary size can improve the test accuracy of the model, up to a certain point. In this study, the maximum test accuracy of approximately 0.83 was achieved for a vocabulary size of 2000 words.

b. Variants comparison

The results presented in Figure 1 demonstrate that the model without the most common words achieves a higher accuracy compared to the other two models. Specifically, the model using the whole vocabulary shows a lower accuracy than the model without the most common words. Additionally, the model with stemming has a lower accuracy than the model without the most common words, but performs better than the model using both techniques.
These findings suggest that removing the most common words from the vocabulary has a positive impact on the accuracy of the model, while stemming can have a slightly negative effect. One possible explanation for these results is that the most common words in the dataset may not carry much semantic meaning and can introduce noise into the model. On the other hand, stemming can potentially remove important distinctions between words and reduce the model's ability to differentiate between them.

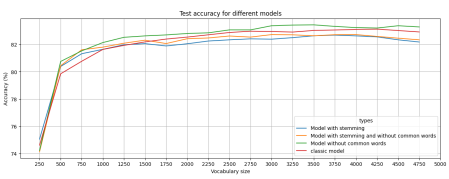
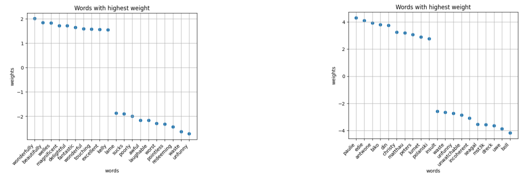


Fig. 1: Comparison of the models for different vocabulary sizes



(a) **(b)**

Fig. 2: Most impactful words, for the model without common words, a- vocabulary size =2000, b- vocabulary size =5000

c. Most impactful words

Figure **??** displays the most impactful words for the model using a 2000-size vocabulary without the most common words. The results show that the model performs well, giving significant importance to adjectives that convey a clear positive or negative sentiment. This outcome can be attributed to the effectiveness of the vocabulary size which enable the model to focus on the most relevant words for classification. However, when the vocabulary size increases to 5000, the model's performance degrades, as seen in Figure **??**. The model appears to give importance to words that do not necessarily carry a clear sentiment, leading to a less effective classification.

d. worst errors on the test set

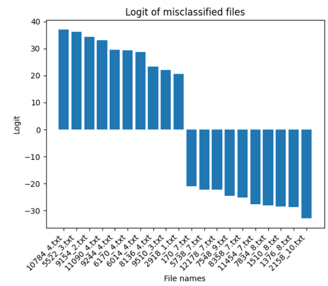


Fig. 3: Logit for misclassified reviews

e. Conclusions