
Movie Reviews: A machine learning project

D. Ligari 518592¹

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

Contact: davide.ligari01@universitadipavia.it

Date: May 8, 2023

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. By using large datasets and advanced algorithms, machine learning models can accurately classify reviews based on their content. In this report, is presented 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.

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. Through 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:

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

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.

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.