Sentiment Analysis on IMDB Reviews

Parand Shams (parandshb@gmail.com)
Department of Computer Science, University of Tabriz, Tabriz, Iran

Abstract

Text classification can be used to categorize documents, reviews, files, and any text from all over the web. In the given task, using an IMDB movie Review, the goal is to determine whether a review is positive or negative. This technique is called sentiment analysis, often performed on textual data to help businesses monitor brands, extract information from customer feedback, and opinion mining. In this paper, several methods were applied, considering that the project was done in two phases, the aim was to compare the models, choose the best one, and enhance the efficacy of the performance. In the first phase, a Multinomial Naïve Bayes classifier (MNB) was used, and in the final phase, a support vector machine (SVM) is applied; each method reached accuracies of 85.5% and 89.8%, respectively.

Keywords: Sentiment Analysis, Text Classification, Opinion Mining, Multinomial Naïve Bayes, Support Vector Machine

1 Introduction

Sentiment analysis of a movie review can help rate how positive and negative a review is, hence viewers' overall opinion regarding a movie. Using a summary of movie reviews, we can decide whether or not to waste time on a movie. This project aims to implement binary classification using several methods to automate understanding the reviews through training, validation, and test sets.

First, we represent a general pipeline for the classification using Multinomial Naive Bayes in phase one. To use the data, we need to preprocess it through the following steps, tokenizing, removing stopwords and punctuation, and lemmatization using the Natural Language Toolkit (NLTK). Then, we need to vectorize each review using TF-IDF, a statistical measure that evaluated the relevance of a word to a document. After that, we need to split the data into a training set and test set to fit and evaluate the presented model. Finally, we predict the test set results and store the results into a confusion matrix. In the second phase, by adding a validation set and implementing several models, we aim to find the best model and boost performance and accuracy. To implement the models, we use Scikit-learn libraries.
The rest of the paper represents the implemented models in section 2, describes and compares the results in section 3, and presents the conclusion in section 4.

2 The Methods

A model uses a training and validation dataset to predict labeled results of given reviews. Through both phases, we use different models to predict the sentiment and label encoding the reviews. Implemented methods are as follows.

2.1 Multinomial Naïve Bayes

Multinomial Naïve Bayes uses term frequency i.e. the number of times a given term appears in a document. Term frequency is often normalized by dividing the raw term frequency by the document length. After normalization, term frequency can be used to compute maximum likelihood estimates based on the training data to estimate the conditional probability.[1]

2.2 Adaptive Boosting

Boosting is a general ensemble method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. Models are added until the training set is predicted perfectly or a maximum number of models are added. AdaBoost, short for ”Adaptive Boosting” was the first really successful boosting algorithm developed for binary classification.[2] [3]

[1] https://www.mygreatlearning.com/blog/multinomial-naive-bayes-explained/

[2] https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/

[3] https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c

2.3 Logistic Regression

Logistic Regression is used when the dependent variable (target) is categorical.It's an extension of the linear regression model for classification problems.[4] [5]

2.4 Random Forest

Random forests or random decision forests are an ensemble learning method for classification by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.[6]

2.5 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. [7]

3 Evaluation and Results

By comparing different methods, it can vividly be understood that the C-Support Vector Classifier has the highest accuracy among other models.[8] Nevertheless, previous work on IMDB long reviews using the combination of Multinomial Naive Bayes and Support Vector Machine (NBSVM) model seems promising as they provided better results and reached a state-of-the-art performance level.(Wang and Manning, 2015[2])
Implementation of models has provided the results represented in Table 1. The top 3 methods are in **bold** and the best is also underlined.

Table 1: Results

Method	F1-Score	Accuracy	AUC score.
MNB	**86**	**85.8**	85.8
AdaBoost	81	81.4	81.3
Logistic Regression	**89**	**89**	89
Random Forest	85	85	85
SVM	**90**	**89.8**	89.8

4 Conclusions

In this paper, we compared different models to classify the sentiment of movie reviews and proposed the method with the highest accuracy rate. Lastly, we achieved an accuracy of 86.22% on average. In future works, we would like to implement the NBSVM method on our dataset to compare whether it reaches a new state-of-the-art level among performed methods.

References

[1] B. Lakshmi DeviV. Varaswathi BaiSomula Ramasubbareddy Email authorK. Govinda. *Emerging Research in Data Engineering Systems and Computer Communications.* , 1993.

[2] Sida Wang and Christopher D. Manning. *Baselines and Bigrams: Simple, Good Sentiment and Topic Classification.* , 2015.

[3] Sentiment Analysis,
`https://monkeylearn.com/sentiment-analysis/`

[4] Text Classification,
`https://monkeylearn.com/text-classification/`

[5] Jason Brownlee: Types of Classification in Machine Learning,
`https://machinelearningmastery.com/types-of-classifi`

[4] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[5] https://christophm.github.io/interpretable-ml-book/logistic.html

[6] https://en.wikipedia.org/wiki/Random$_{forest}$

[7] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[8] Regularization parameter equal to 1E5 and kernel type rbf.