Don Castillo CPSC 4660
001224541 January 28, 2022

Movie Review Classifier Using Naïve Bayes Classification Strategy

Introduction

Reading reviews on websites like IMDB, Yahoo! Movies, and Rotten Tomatoes is one way of determining whether it is worthy of time to watch a movie or not. Movie reviews represent the feeling of the audience toward the movie - they could be categorized as good or bad. Usually when writing a movie review, not only the reviewers provide an in-depth analysis of the movie, they also provide a rating, a numerical value that represent their overall feeling about the movie. In a rating scale of 1 through 5, a rating of 1 could mean bad review, 3 could mean neutral, and 5 could mean good review. The website then averages the ratings of all the reviewers to come up with a single number that represent the overall sentiment of all the reviewers. An average rating of 4.5 could mean that the movie is generally good.

Naïve Bayes Classifier

What if we remove the rating scale and leave the text to represent a movie review? How do we now know if the review is good or bad? How do we classify a text? This is what the Naïve Bayes Classifier can do. Naïve Bayes is a classification strategy that predicts the category of the text based on the words (called features) in it, the frequency of the words, and how the words were used. It is a supervised machine learning classifier - meaning the classifier will have to be trained with a set of preclassified texts or document. Once it learns the features associated with a category from this training document, it can predict what class should a text belong to.

Naïve Bayes Classification has been a popular subject in many machine learning studies. It is one of the classification methods used by Jain et al. (2021) in their study where they compared and analyzed different fake news classifiers.

The Project and Implementation Workflow

For the CPSC 4660 Project, the programming project will make use of Naïve Bayes Classification strategy to implement a simple Node.js-based movie review classification web app. The idea is to predict the movie review entered by the user.

Firstly, the data will be coming from an open-source movie review dataset (Maas, et al., 2011) based on past reviews on IMDb. It is available on http://ai.stanford.edu/~amaas/data/sentiment/. The dataset will be processed (i.e. tokenize the words, remove stop words, and count the frequencies of the words). The raw data and processed data will be stored in a database. MongoDB will be used as the database to achieve speed and flexibility in processing text.

Secondly, the dataset will be split into training and testing data. All data in the dataset is preclassified, meaning they are already classified as positive or negative. A Naïve Bayes Classification model containing the algorithm to train and test will be implemented as suggested by (Jurafsky & Martin, 2020).

Using the training data, the model will be evaluated using four evaluation metrics. These evaluation metrics are accuracy (ratio of correct predictions to the total number of the input data), precision (ratio of many events correctly remembered to the number of all events remembered), recall (percentage of items actually present in the input that were correctly identified by the system), and the f measure (harmonic mean of precision and recall). The work of Jain et al. (2021) with the explanation provided by Jurafsky and Martin (2020) will be used as a guide to evaluate the implemented Naïve Bayes Classification model for the movie review app.

References

Jain, R., Jain, D. K., Sharma, D., & Sharma, N. (2021, December). Fake News Classification: A Quantitative Research Description. *ACM Trans. Asian Low-Resour. Lang. Inf. Process., 21*.

Jurafsky, D., & Martin, J. H. (2020). Naive Bayes and Sentiment Classification. In D. Jurafsky, & J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing,* (3rd ed., pp. 55-73).

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142-150. Retrieved from https://aclanthology.org/P11-1015/