Movie Review Sentiment Analysis

Hanlin Zhang, Jingyi Zhou, Qishun Wang

April 5, 2019

Problem statement Sentiment analysis is a major subject in NLP realm. It is also a sub-domain of opinion mining where the analysis is focused on the extraction of emotions and opinions of the people towards a particular topic from a structured, semi-structured or unstructured textual data[2]. Reviews provide us with a qualitative way to gauge the performance of a movie while ratings tell us the how successful a movie is quantitatively. Analyzing the relationship between reviews and ratings can help us understand how a movie generally meets the expectations of its reviewers.

The data The dataset we plan to use is a collection of movie reviews retrieved from the Rotten Tomatoes website by Pang and Lee[1], which is publicly available on a Kaggle competition. To set the benchmark, the original dataset has been split into train/test. Since the test contains no labels, we are going to use the the train as our project dataset. The labeled dataset consists of 156k movie reviews along with their phraseIDs and sentenceIDs. To be specific, it is a multi-class dataset labeled on a scale of 0(negative) to 4(positive), which presents different levels of sentiment.

The model In the original dataset, each sentence has been parsed into many phrases by the Stanford parser, and each phrase is labelled by its own sentiment. The labels are not just positive or negative but on a multi-point scale (e.g., 0 to 4), which also makes this problem more challenging. We will begin with re-weighting the data using tf-idf or bag of words, and then start from multiclass logistic regression using strategies like one-vs-all and multinomial. Linear SVM is another model we want to implement, and we will also use different strategies for multiclass classification. At last, we will try to implement a neural network such as 1D CNN. Neural network with multiple layers should give us a good result as long as we have enough computing power.

How the model will be evaluated It is a classification problem, so we will choose accuracy, precision and F1-score as metrics of performance, as well as the original test and training set accuracy. And we will implement cross validation since our dataset is fairly large enough to be separated to several folds.

Anticipated challenges Anticipated challenges are 1) finding the optimal solution for tokenization the documents; 2) identifying a group of promising classification algorithms that works best for sentiment analysis ; 3) facing a lack of computational power if applying neural networks. We will try to overcome these difficulties by conducing comparison and contrast of the performance among different algorithms and doing multiple experiments with different tokenization strategies. To ensure the best outcome and avoid over-fitting, we are also going to do grilled search and cross validation for the best parameters.

Data visualization Data visualization is essential to the problem we are looking at and it provides insight that can be easily understood without diving into the math. We are going to create multiple data visualizations in the final report with existing tools that are available in python libraries (e.g., matplotlib, seaborn, D3).

References

[1] Pang Bo and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL*, pages 115–124, 2005.

[2] Tirath Prasad Sahu and Sanjeev Ahuja. Sentiment analysis of movie reviews: A study on feature selection and classification algorithms. *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, pages 1–6, 2016.