MOVIE SENTIMENT ANALYSIS

David Fu[1], Cynthia Loh[1]
Aakash Irengbam, Aniket Patel

Stevens Institute of Technology
dfu6@stevens.edu, cloh@stevens.edu,
airengba@stevens.edu, apate211@stevens.edu

ABSTRACT

1 Introduction

There has been a rapid, notable increase of movie reviews written by the viewers. This can be credited to websites, such as IMDB, that encourage users to write feedback and reviews. Movie reviews are generally a great way for potential viewers to gauge which movie they would enjoy watching or not; however, the sheer amount of reviews make it extremely difficult and unreasonable for a person to read through completely. In order to overcome this challenge, a way to automate movie review classification and summarize the average sentiment of a movie is necessary. This would allow customers to accurately and quickly comprehend the average sentiment of a particular movie without any exhaustive effort. The objective of this project is to perform a sentiment analysis on a given dataset of highly polarising movie reviews. So we classify the reviews as either positive or negative. The dataset we are using contains movie reviews gathered from IMDB as well as their binary sentiment classification of either positive or negative. In total there are fifty-thousand movie reviews and the dataset is split evenly for training and testing data. In the entire dataset, there are no more than 30 reviews from the same movie because reviews for the same movie typically have correlated ratings.

2 Data Analysis

Initial analysis of movie synopsis the data have been classified into positive and negative sentiment. Along with the given vocabulary, the data represents comprehensive all the words that were used consist of ninety thousand feature words. With some pre-processing and optimization the number will be reduced significantly.

For example a sample dataset looks like the following.[1]

”Fair drama/love story movie that focuses on the lives of blue collar people finding new life thru new love.The acting here is good but the film fails in cinematography,screenplay,directing and editing.The story/script is only average at best.This film will be enjoyed by Fonda and De Niro fans and by people who love middle age love stories where in the coartship is on a more wiser and cautious level.It would also be interesting for people who are interested on the subject matter regarding illiteracy..”

This is an example for positive sentiment. Upon confirmation and reading, most reviews are classified pretty accurately.

3 Method

Using term frequency-inverse document frequency (TFIDF), we were able to extract word features from the dataset. This results in extracting many words that are articles, prepositions, pronouns, and conjunctions. These words, also known as stopwords, are usually the most common words in the English language and do not provide any relevant or important information to the sentiment analysis of the reviews. In order to optimize for performance we remove

these stopwords. To further improve performance, we disregard complex, hyphenated words as well. Although performance would improve, this may substantially reduce the accuracy due to lack of specialized features. After data-preprocessing, we can perform binary sentiment classification using machine learning algorithms. There are five algorithms that will be used to classify the movie reviews as either positive or negative. The algorithms to be used for this project: knn, neural networks, naive bayes, logistic regression, Long Term Short Term Memory (LTSM).

References

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis, June 2011.