SENTIMENT ANALYSIS OF MOVIE REVIEWS

Group No. 2

Anveshak Rathore	181060012
Darshan Kedari	181060021
Vedant Mankar	181060040
Dixit Mendon	181060043

AIM OF THE PROJECT

To predict the number of positive and negative reviews using sentiment analysis technique on the review corpus and a suitable classification model.

INTRODUCTION

Movie reviews are an important way to gauge the performance of a movie. While providing a numerical/stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews is what gives us a deeper qualitative insight on different aspects of the movie. A textual movie review tells us about the strong and weak points of the movie and deeper analysis of a movie review can tell us if the movie in general meets the expectations of the reviewer.

SENTIMENT ANALYSIS

Sentiment Analysis is a major subject in machine learning which aims to extract subjective information from the textual reviews. The field of sentiment of analysis is closely tied to natural language processing and text mining. It can be used to determine the attitude of the reviewer with respect to various topics or the overall polarity of review. Using sentiment analysis, we can find the state of mind of the reviewer while providing the review and understand if the person was “*happy*”, “*sad*”, “*angry*” and so on.

In this project we aim to use Sentiment Analysis on a set of movie reviews given by reviewers and try to understand what their overall reaction to the movie was, i.e. if they liked the movie or they hated it. We aim to utilize the relationships of the words in the review to predict the overall polarity of the review.

DATASET

The dataset used for this task was collected from *Large Movie Review Dataset* which was used by the AI department of Stanford University for the associated publication. The dataset contains 50,000 training examples collected from IMDb where each review is labelled with the rating of the movie on scale of 1-10.

We then tried to redistribute the examples as 40,000 for training and 10,000 for testing.

PRE-PROCESSING TASKS

- ★ One necessary pre-processing step prior to feature extraction was removal of HTML tags like “<br>”. We used simple regular expressions matching to remove these HTML tags from the text.
- ★ Another important step was to make the text case-insensitive as that would help us count the word occurrences across all reviews and prune unimportant words.
- ★ We also removed all the punctuation marks like ‘!’, ‘?’, etc. as they do not provide any substantial information and are used by different people with varying connotations.
- ★ We also removed stopwords from the text for some of our feature extraction tasks.

FEATURE EXTRACTION

We used 2 methods for extraction of meaningful features from the review text which could be used for training purposes. These features were then used for training several classifiers.

- **Bag of Words**: This is a typical way for word representation in any text mining process. We first calculated the total word counts for each word across all the review and then used this data to create different feature representations. As the total number of words in the dictionary was huge more than 1,60,000) the first feature set was created using only the 50,000 most frequent words according to their occurrence

- **TF-IDF Modelling:** While the method of feature extraction described above concentrated more on higher frequency parts of the review they completely ignored the portions which might be less frequent but have more significance for the overall polarity of the review. To account for this, we created feature representations of words using *TF-IDF*. The feature representation for this model is similar to the Bag of Words model except that we used TF-IDF values for each word instead of their frequency counts.

MODEL USED

For this classification task we explored multiple classification models   on above feature representations. However, we obtained the best results with the simple **Logistic Regression** model.

ACHIEVABLE OUTCOMES ON IMPLEMENTATION

- Obtains a one-glance review derived from the many customer reviews that are available
- Reduces the strain of going through multiple unfiltered word reviews, and gives more reliable result than a star-based review system
- Determines the most frequently used words while expressing a positive or negative sentiment

FUTURE IMPROVEMENT PROSPECTS

One of the major improvements that can be incorporated as we move ahead in this project is to merge words with similar meanings before training the classifiers. Another point of improvement can be to model this problem as a multi-class classification problem where we classify the sentiments of reviewer in more than binary fashion like “Happy”, “Bored”, “Afraid”, etc.

This problem can be further remodeled as a regression problem where we can predict the degree of affinity for the movie instead of complete like/dislike.

CONCLUSION

For the movie review analysis problem, we have successfully implemented a Logistic Regression model, along with suitable data pre-processing, thus yielding good accuracy results.