RATINGS PREDICTION

Submitted by:

GOURAV KUMAR

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

INTRODUCTION

- The rise in E — commerce, has brought a significant rise in the importance of customer reviews. There are hundreds of review sites online and massive amounts of reviews for every product. Customers have changed their way of shopping and according to a recent survey, 70 percent of customers say that they use rating filters to filter out low rated items in their searches.

- The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is vital to companies that support these reviews, companies like Google, Amazon and Yelp!.

- There are two main methods to approach this problem. The first one is based on review text content analysis and uses the principles of natural language process (the NLP method). This method lacks the insights that can be drawn from the relationship between costumers and items. The second one is based on recommender systems, specifically on collaborative filtering, and focuses on the reviewer's point of view. Use of the user's similarity matrix and applying neighbors analysis are all part of this method. This method ignores any information from the review text content analysis.

- In an effort to obtain more information and to improve the prediction of the review rating, the researchers in this article proposed a framework combining review text content with previous user's similarity matrix analysis. They then did some experiments on two movie review datasets to examine the efficiency of their hypothesis. The results that they got showed that their framework indeed improved prediction of the review

rating. This article will describe my attempt of following the work done in their research through examples from the Amazon reviews dataset. The notebook documenting this work is available here and I encourage running the code on your computer and report the results.

Model/s Development and Evaluation

In order to check and choose the best model, I constructed a pipeline that did the following steps. The pipeline will first perform a TF-IDF term weighting and vectorizing and will then run the classification algorithm. In general, TF-IDF will process the text using my “text_process” function from above, and then convert the processed text to a count vector. Afterwards, it will apply a calculation that will assign a higher weight to words of more importance.

Note that I chose ngram_range = (1, 2) and that the algorithm was Multinomial Naïve Bayes. Those decisions were taken according to the results of a cross-validation test. The cross-validation test that I did is beyond the scope of this article, but you can find it in the notebook.

The models checked were:

1. Multinomial logistic regression, as a benchmark
2. Multinomial Naïve Bayes
3. Decision Tree
4. Random forest

Multinomial Naïve Bayes gave the best accuracy score[1] (0.61) and therefore the predictions made by it were chosen to represent the RRP based on RTC.

CONCLUSION

- In conclusion, my thesis was proven to be correct. Combining the formerly known data about each user's similarity to other users with the sentiment analysis of the review text itself, does help improve the model prediction of what rate the user's review will get
- This paper goal is to compare the methods and to see if the framework offered by the researchers will improve the predictions accuracy. It was not to find the most accurate model for RRP based on RTC.
- Although MAE of 0.66 is not good, the main aim of this work was to check the hypothesis and not necessarily to seek the best RRP model.