Using Sentiment Analysis to Determine the Average in a 1-10 Scoring System

Omar Morales
Computer Science Department
Humboldt State University
Arcata, CA USA
om61@humboldt.edu

Sherrene Bogle
Computer Science Department
Humboldt State University
Arcata, CA USA
sherrene.bogle@humboldt.edu

***Abstract*—This paper examines the 10-star rating systems used in movie reviews and looks at the consistency between reviews. This paper also explores the idea of an average score and what section of a 10-star represents positive ratings versus negative ratings. A support vector machine is used to build a model and classify reviews based on scores and sentiment. After the accuracy and absolute errors are determined for each score, all scores are compared with one another. The support vector machine that was used for classification had an accuracy of 86.48% and an error rate of 0.371. After comparing the different star reviews, the best representation for an average score, as described in the paper, was the 6-star review. This is because 50.72% of the reviews were classified as positive with an error rate of 0.496, meaning half were considered to have a positive sentiment and half were considered to have a negative sentiment.**

Keywords—machine learning, sentiment Analysis, movie reviews, scoring systems, support vector machine, SVM

I. INTRODUCTION

The goal of this paper is to study the distribution of scores on a 10-star scoring system One of the most commonly used scoring systems used to evaluate different forms of media is the 1-10 scale, with 1 being the lowest rating and 10 being the highest. In this paper one of the questions being asked is," what is considered an average score?". This question, however, was not as straight forward as initially though. This is because the idea of average in the everyday lexicon is not exact. Based on normal convention, it would stand to reason that a score of 5.5 would represent an "average" score, considering that 5.5 is the arithmetic mean of ten. Although this may be true, the word average, when in reference to a review, can also be used to represent a score that is in between a positive review and a negative review. In this case any score above this "average" score would be considered positive and anything below would be considered negative. Sine the 10-star scale that is being studied is designed to represent the full range of reviews, from terrible to excellent, a case can be made that a 5.5 score could be considered an "average" score by both definitions mentioned.

In this study, the difference between the two different definitions of average were examined and compared with one another. In addition to this the researchers of this paper decided to analyze IMDB movie reviews to test the consistency between the score that was given and the sentiment of the review itself. By comparing the score that was given with the overall sentiment of the review, a conclusion can be drawn about the internal consistence of reviews and whether a score of 5.5 is indeed an average score. This study could help bring context to the scores that movies receive, which could be of great use to the studios that produce movies. Likewise, this study could help expose the patterns present in online movie reviews and expose bias or common traits found in these reviews.

II. LIT REVIEW

A. Sentiment Classification Of Movie Reviews By Supervised Machine Learning Approaches

Other studies have looked at the accuracy of using a SVM model for sentiment analysis on movie reviews. In a 2013 study by P. Kalaivani and Dr. K.L. Shunmuganathan [1], a SVM was used to classify movie reviews based on their sentiment. Along with an SVM, other machine learning algorithms were used in order to build models for sentiment analysis. The study showed that an SVM based model can be trained reach more than 80 % of the classification correctly with as little as 1000 total reviews.

B. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales

The star rating system in relation to sentiment analysis has be studied such as in the 2005 study conducted by Bo Pang and Lillian Lee [2]. In this study Pang and Lee used different supervised classification methods to determine the opinion polarity of a given document. Pang and Lee also examine a human's ability to distinguish between reviews of different star ratings. Their results, although limited, show that humans are able to accurately determine the star rating of a review on a 5-star scale. This experiment differs from the one in this paper because it studies the 5-star rating system as opposed to the 10-star system, however it does reinforce the idea that there are discernible qualities present in reviews that different them from reviews of different ratings.

C. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews

In a paper publish by Kushal Dave, Steve Lawrence, and David M. Pennock [3], product reviews were examined and labeled based on their polarity. The reviews that were studied used a 5-

star rating system similar to the work done by Pang and Lee [2]. This paper provides some very useful insight when it comes to the problem with reviews left by people on the web. One of the issues that arose during the research of Dave, Lawrence and Pennock, was the inconsistency with the rating that was given and the language that was used in the review itself. The paper explains that a portion of the reviews contain very negative language throughout the review. The only positive section of the review being the final sentence in which the reviewer expresses that they are satisfied with the product. This aversion to giving a negative review may help explain some of the results that were found during this paper's research.

D. Sentiment Regression: Using Real-Valued Scores to Summarize Overall Document Sentiment

Research on rating systems with higher gradation has also been done in the past. In the 2008 study done by Adam Drake, Eric Ringger, Dan Ventura [4], a 91-point scale is studied. In this study a scale that ranges from 1-10 is used. Unlike the 1-10 scale that is being studied in this paper, the scale in the 2008 paper [3] breaks down the 1-10 scale into increments of 0.1 starting at 1.0 and ending at 10.0. The results of the experiment described in the paper indicate that it is possible to accurately classify scores within 1.21 of the true score.

E. Multiple Aspect Ranking using the Good Grief Algorithm

There are several issues that arise when trying to classify reviews based on sentiment. Benjamin Snyder and Regina Barzilay explain where some error might occur in the classification when the review is not dived into its parts [5]. In their study, Snyder and Barzilay examine different aspects of food reviews on a 5-star scale. The study indicated that a classification model preforms better when the reviews are subdivided into multiple parts based on what aspect they are reviewing. For example, a user might rate the food of a restaurant very high while rating the service of wait staff very low. These different aspects of the overall experience have different weights in the eyes of the reviewers which may affect the final score that is given to the restaurant. This idea of breaking down reviews into different parts could have helped increase the accuracy of the results found in this very paper, however no such techniques were used.

III. METHODOLOGY

The data set that is being used to train the model used in this study comes from the 2011 ACL paper titled *Learning Word Vectors for Sentiment Analysis*. The data set consists of 50,000 review evenly divided between positive and negative reviews. The reviews are also evenly divided into a learning and training dataset, each consisting of 25,000 reviews each. In this paper the 25,000 reviews that are present in the training dataset were used. The negative reviews include score ranging from 1-4 stars. The positive reviews consist of scores ranging from 7-10 stars. No review of 5 stars or 6 stars are included in the training set. In order to collect the movie reviews needed to test the model, a data scraper was used to harvest reviews from various movies on the IMDB website. The data scraper that was used in this study was a web extension for the Google Chrome Web browser named *Instant Data Scraper* created by *webrobots.io*. The data that was collect consists of 5,820 reviews that range from 4 to 6 stars. The reviews that were collected are evenly split between the three ratings resulting in 1,940 reviews each. As for the rest of the reviews, scores that ranged from 1-3 and 7-10 were collected from the testing dataset that are included in the IMDB dataset that was described earlier. Overall, reviews were divided into individual Excel files based on their rating, 1-10. Each rating has 1940 reviews attached to it in order to create a testing dataset that consists of a total of 19,400 reviews.

The 4-star scores and 5-star scores were labeled as negative reviews and the 6-star reviews were labeled as positive. This was done because, as explained in the introduction to this paper, a score of 5.5 is being assumed to be an average score. Due to the nature of the rating system being studied, fractions of stars are not allowed. Due to this limitation, any score equal or less than 5-stars will be considered negative and any score equal to or greater than 6-stars will be considered positive. Even though this compromise is necessary it does not directly violate the definition of average that was described in the introduction to this paper. This is because a score of six still lies above a score of 5.5, therefore it can be considered positive. Likewise, a score of five lies below 5.5 so it can be considered negative.

A. Data Preparation

Originally the data set was comprised of reviews which were in individual text files. The step that was taken in the data preparation step was condensing the data into a single Excel file which had two attributes. The first attribute was the sentiment of the, either negative or positive. The second attribute was the text of the review itself. In addition to this a process known as tokenization was used. Tokenization is the process of separating a larger piece of data into individual tokens. In this case, the movie reviews were broken down into their individual words. All tokens were then transformed to uppercase and stop words were filtered out. In this context, stop words refers to common English words such as "and", "you", and "the".

B. Experimental Setup

- The model used in this paper to classify reviews based on their sentiment was created using a Support Vector Machine (SVM). The model that was created was able to accurately classify reviews based on their sentiment with an accuracy of 88.25% and an absolute error of 0.371(TABLE I).
- Ten K-fold cross validation was used during the training of the model. The SVM classifier is, by design, resistant to overfitting, therefore no additional steps were taken during the training on the model in order to minimize overfitting.
- Each rating, 1-10, has an individual Excel file with the reviews that accompany it. Each file is tested individually in order to compare each score independently.
- Scores that ranged from 1-5 were labeled as negative while score between 6-10 were labeled as positive.

IV. RESULTS AND ANALYSIS

After running the SVM model on all the different star ratings some trends did appear. Fig 1. shows that the accuracy of classification tended to be greatest at the ends for the graphs and gradually got less accurate as scores approached the score of six. The accuracy was the highest for 1-star reviews at 87.53% and steadily deceased as previously stated. The accuracy continued to trend downward until it reached its lowest point of 50.72% at the 6-star rating. Scores following the 6-star rating trended upwards until the reached the 10-star score which had an accuracy of 88.25%. The inverse was true for the absolute error as shown in Fig. 2. The lowest error rate was lowest for the 1-star reviews and increased every rating until it reached its maximum at the 6-star reviews with an absolute error of 0.496. After the 6-star reviews, the absolute error decreased until it reached the second lowest error rate of the whole graph with the 10-star reviews with an absolute error of 0.368. In both cases the score with the best correct classification rate was the 1-star reviews with an accuracy of 87.53% and an absolute error of 0.362. The next best was the 10-star rating with an accuracy of 88.25% and an absolute error of 0.368.

Both Fig.1 and Fig. 2 show that the model struggled most with the 6-star ratings. When it came to accuracy, the 6-star reviews were correctly predicted only 50.72% of the time which is effectively as accurate as a coin toss. The 6-star reviews were also 11.39% less accurately classified as the second worst preforming rating which was the 5-star ratings at 62.11%. The same pattern occurs when considering the absolute error. 6-star reviews again preformed the worst with the highest error rate of any other score with an absolute error of 0.496. The 6-star review is the obvious outlier in the results. No other reviews showed such a dramatic change between scores. In every other case, the accuracy only changed by about 8% on average. For example, the difference in accuracy from 1-stars to 2-stars is only 6.1% or in the case of 8-stars to 9-stars, the difference is only 1.3%. When it comes to the 6-star reviews and its neighbors, the difference was at most 22.7% percent, when compared to the 7-star reviews.

TABLE I

SVM Model			
	True Positive	***True Negative***	***Prediction***
pred. positive	11377	2257	83.45%
pred. negative	1123	10243	90.12%
class recall	91.02%	81.94%	
Acurracy	86.48%		
Absolute error	0.371		

TABLE II

1-Star Reviews			
	True Negative	***True Positive***	***Prediction***
pred. negative	1698	0	100.00%
pred. positive	242	0	0.00%
class recall	87.53%	0.00%	
Acurracy	87.53%		
Absolute error	0.362		

TABLE III

2-Star Reviews			
	True Negative	***True Positive***	***Prediction***
pred. Negative	1580	0	100.00%
pred. Positive	360	0	0.00%
class recall	81.44%	0.00%	
Accuracy	81.44%		
Absolute error	0.388		

TABLE IV

3-Star Reviews			
	True Negative	***True Positive***	***Prediction***
pred. Negative	1505	0	100.00%
pred. Positive	435	0	0.00%
class recall	77.58%	0.00%	
Accuracy	77.58%		
Absolute error	0.415		

TABLE V

4-Star Reviews			
	True Negative	***True Positive***	***Prediction***
pred. Negative	1356	0	100.00%
pred. Positive	584	0	0.00%
class recall	69.90%	0.00%	
Accuracy	69.90%		
Absolute error	0.437		

TABLE VI

5-Star Reviews			
	True Negative	***True Positive***	***Prediction***
pred. Negative	1205	0	100.00%
pred. Positive	735	0	0.00%
class recall	62.11%	0.00%	
Accuracy	62.11%		
Absolute error	0.461		

TABLE VII

6-Star Reviews			
	True Positive	***True Negative***	***Prediction***
pred. Positive	984	0	100.00%
pred. Negative	956	0	0.00%
class recall	50.72%	0.00%	
Accuracy	50.72%		
Absolute error	0.496		

TABLE VIII

7-Star Reviews			
	True Positive	***True Negative***	***Prediction***
pred. Positive	1425	0	100.00%
pred. Negative	515	0	0.00%
class recall	73.45%	0.00%	
Accuracy	73.45%		
Absolute error	0.428		

TABLE IX

8-Star Reviews			
	True Positive	***True Negative***	***Prediction***
pred. Positive	1637	0	100.00%
pred. Negative	303	0	0.00%
class recall	84.38%	0.00%	
Accuracy	84.38%		
Absolute error	0.389		

TABLE X

9-Star Reviews			
	True Positive	***True Negative***	***Prediction***
pred. Positive	1661	0	100.00%
pred. Negative	279	0	0.00%
class recall	85.62%	0.00%	
Accuracy	85.62%		
Absolute error	0.381		

TABLE XI

10-Star Reviews			
	True Positive	***True Negative***	***Prediction***
pred. Positive	1712	0	100.00%
pred. Negative	228	0	0.00%
class recall	88.25%	0.00%	
Accuracy	88.25%		
Absolute error	0.368		

FIGURE I

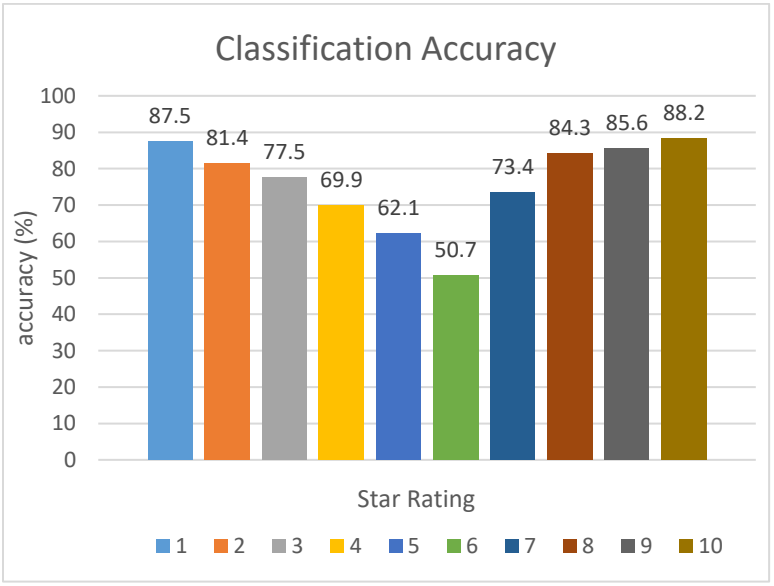
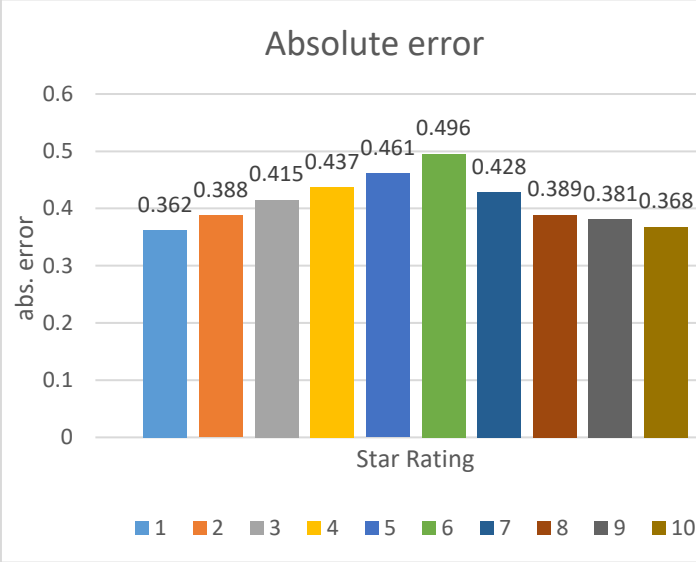


FIGURE II



V. CONCLUSION

After examining the results of the study, it was found that the SVM model that was built struggled the most when it came to classify the 6-star reviews. Fig. 1 shows that the accuracy of the 6-star reviews were by far the worst. Considering that 5 and 6-star reviews were absent during the initial training of the model, it makes sense that these two ratings performed the worse overall. What is surprising, on the other hand is that the 6-star reviews preferred the worst. Based on initial predictions, the score of 5-stars should have preformed closer to the 50% accuracy mark. This is because, based on the assumption made at the beginning, a score of 5-stars would be considered an average score. This, however, more accurately describes the results that were found with the 6-star reviews. The SVM model classified 984 of the reviews as positive and 956 reviews as negative. This could suggest that the 6-star review is a better representation of an average score when it comes to online movie reviews. This would mean that any score equal to or greater than a seven would be considered positive and any score equal to or than a five would be considered negative.

Future research could be done exploring this hypothesis and testing the validity of the results found in this paper. A study comparing the results of this paper with human participants could prove to be useful since it has shown that humans are more perceptive to small details in reviews that would help differentiate them based on score [2]. Separating each review into its different components could also help improve the accuracy of the model and could be another area of study worth considering.

ACKNOWLEDGMENT

The data set used is collection of 50,000 IMDB reviews gathered from various movies. The data set comes from a paper titled, *Learning Word Vectors for Sentiment Analysis* written by Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, and Christopher.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).

REFERENCES

[1] Kalaivani, P., & Shunmuganathan, K. L. (2013). Sentiment classification of movie reviews by supervised machine learning approaches. Indian Journal of Computer Science and Engineering, 4(4), 285-292.

[2] Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075.

[3] Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

[4] Drake, A., Ringger, E., & Ventura, D. (2008, August). Sentiment regression: Using real-valued scores to summarize overall document sentiment. In 2008 IEEE International Conference on Semantic Computing (pp. 152-157). IEEE.

[5] Snyder, B., & Barzilay, R. (2007, April). Multiple aspect ranking using the good grief algorithm. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (pp. 300-307).