Predicting Individual Star Ratings From Reviews on Yelp

Laree Block*, David Haehlen*, Alexander Krusina*, Mohammad Noaeen*, Abdullah Sarhan†, Zahra Shakeri Hossein Abad‡
*Department of Electrical and Computer Engineering, University of Calgary, Canada
{lpblock, dhaehlen, alexander.krusina, mohammad.noaeen} @ucalgary.ca
† Department of Computer Science, University of Calgary, Canada, asarhan@ucalgary.ca
‡ Department of Community Health Sciences, University of Calgary, Canada, zshakeri@ucalgary.ca

***Abstract*—Yelp is a website that allows users to review a business, it's products, and it's services. There are reams of reviews for businesses in Yelp's database, most consisting of a star rating and a free text review. These reviews are used by potential customers to find new businesses and compare options. We have explored these reviews available in Yelp's Open Dataset by using supervised data analysis methods. We have examined the relation between the text of a review and the star rating given to determine if a star rating can be predicted from the text alone. We were able to predict the star rating with an overall accuracy of 63.28%. When categorizing the reviews as positive or negative, we were able to classify the reviews with an accuracy of 86.70%. Additionally, have determined that the location of a business (latitude and longitude) and the total number of a reviews a business has some influence on the star rating given.**

***Index Terms*—Yelp review, star rating, natural language processing, machine learning, logistic regression, Naive Bayes, random forest**

I. Introduction and Related Work

Many people rely on reviews and the star rating of various business such as restaurants, bars, and doctors, in order to choose the best option. Most of the time people will have an initial look at the star rating and if they deem it to be too low they will move on to the next option. If the business passes this initial test people may be more intrigued to look at the reviews but some may not go this far. Therefore, the star rating is probably the most important aspect of a review in order to attract new customers through Yelp, where an extra half-star rating can contribute to an increase of 19 % in sales of restaurants [1].

Yelp is one of the most widely used websites to review businesses. Since 2005 Yelp has had over 177 million reviews [2]. Yelp's Open Dataset has been analyzed by many researchers. Ganu et al. [3] put out a study exploring sentiment analysis of text reviews and whether there was an ability to predict ratings from that information. They showed that sentiment analysis of text reviews can help better predict the star rating given for a particular review. Their idea has been widely applied to Yelp reviews. Jong [4] and Xu et al. [5] showed text reviews could be classified into positive and negative reviews fairly accurately. Jong [4] achieved around 73% accuracy with an SVM and 78% accuracy with a Naive Bayes classifier [4]. Various machine learning techniques and preprocessing steps were taken by each team, however, all the results showed that a positive/negative classification was approachable. Some researchers built on this and aimed to predict the exact star rating that a review would give [6]–[8]. Elkouri was able to achieve 63.92%, showing that with machine learning a star rating could be predicted better than with random chance [8]. Other works such as [9] show similar results. In all these papers, only the text review, its sentiment analysis, and the star rating are considered to predict the star rating. Most of these works propose, for future work, adding additional factors that are available in the Yelp Dataset as potential to improve the predictive models performance.

In this paper, we will discuss how a sample of these reviews are analyzed and used to predict the associated star rating. We will also address if there are any patterns in the language used in the review and the star rating to see if it is common that different people consider relatively similar experiences to be a different star rating.

We have used various machine learning techniques including Natural Language Processing (NLP) to help us in predicting the star rating. We have also identified which factors are most important when determining the star rating other than the textual comment. Some of the factors we have looked into include the business location, overall rating, and the number of people that think the review is helpful. We believe that these other factors would allow for a more accurate prediction of the star rating.

The remainder of the this paper is structured as follows: Section II will provide the questions we are answering and how we have processed and analyzed the data. Section III will discuss what results that we have obtained. Section IV will address the results and limitations of this study, and potential future work. Finally, section V will conclude this paper.

II. Methodology

A. Research Questions

This research aims at addressing the following research questions (RQs).

RQ1: *Is it possible to predict a customer's star rating on Yelp from their comments?*
This question was the main topic of our research. It aims to explore how accurate the language in a review is with regards to the star rating. We also wanted to see if it is possible to predict the star rating accurately or if the people leaving the reviews do not use enough common language to do so.
RQ2: *What factors, other than the sentiment of a user's textual review, can be used to predict their star rating?*
This question hopes to explore the various other factors used on Yelp and their relation to the user's individual star rating.

B. Data Collection and Preparation Plan

For the purpose of this study a large amount of Yelp reviews have been gathered. This data includes the user's comments, their star rating, the category and type of business, etc. After obtaining the data and uploading it to Spark, the reviews were processed using the following procedure:

1. Removing non-English reviews: Having non-English reviews will create issues when training the models.
2. Conversion to lower case: Having all of the words in lower case will eliminate the possibility for the word to be considered multiple times.
3. Removing punctuation: This is considered good practice and an important part of text analytics.
4. Removing stop words: Stop words are connecting words that typically add little to no meaning to the text. A sample of common stop words is as follows: to, the, of, is, be, and or. It is crucial that these words are removed as they greatly change the overall word frequencies which will affect the results of the classifiers.
5. Stripping white spaces: This is done in order to ensure the extra white spaces are not counted as additional words.
6. Stemming: Completed by removing the excess characters representing plurality and tenses. Bringing the word back to it's root.
6. Splitting into Categories: This is done because reviews for different types of businesses will typically have different meaningful words. For example, the word "hair" would likely be a stop word in a review for a salon, but would have a negative meaning for restaurants.
7. Manual transformation: This is the final step where custom stop words are removed for each category.

C. Data Analysis

In order to split the reviews into different categories, a list of the most common categories were created. From there, any reviews with multiple categories were narrowed down to the most common category of all they had listed. The list of most common categories was then remade, and the top four categories were chosen to perform an analysis on. The four categories (in order of most common to least common) are:

1. Restaurants
2. Shopping
3. Beauty and Spas
4. Event Planning and Services

The next step of the analysis ties into the final step of the pre-processing. A list of the most common unigrams for each category was created. From this, a list of custom stop words was created by manually browsing through the list and picking words that added no context to the quality of the review. "Food" is an example of a stop word for restaurants. Upon completion of the pre-processing, four different NLP classification machine learning techniques were completed for each of the categories. The techniques used are:

1. Logistic Regression
2. Logistic Regression with TF-IDF (Term frequency - inverse document frequency)
3. Naive Bayes
4. Random Forest

The first results were found using standard values for the parameters for each algorithm. After the initial findings, the parameters for each method were tweaked many times in order to obtain as high of an accuracy as possible, while avoiding using values that would lead to over-fitting.

III. Results

By completing this project, we have determined that it is possible to predict star ratings almost four times higher than randomly guessing, with our highest accuracy being 79.22% (Logistic regression for the Beauty and Spa category). Across all categories, logistic regression with and without TF-IDF performed closely in regards to the accuracy, f-score, and the AUC (area under curve) measures. However, the models without TF-IDF slightly outperformed the models with it. This is likely due to the fact that the majority of the documents (text reviews) are small. Naive Bayes performed similar to logistic regression, the accuracy was 0-2% lower, though the f-score was about 3-5% higher. Random forest was the worst classifier for all categories. It had the worst accuracies, f-scores and AUCs. It also took significantly longer to train the model. The accuracies and f-scores can be seen below.

Naive Bayes was consistently the fastest model to train, which is another reason to choose it over the other classifiers. Logistic regression took 7%-138% longer than Naive Bayes, with the large percent difference on the smaller categories.

Logistic regression with TF-IDF took 22%-82% longer than Naive Bayes, and random forest took 1167%-1179% longer than Naive Bayes.

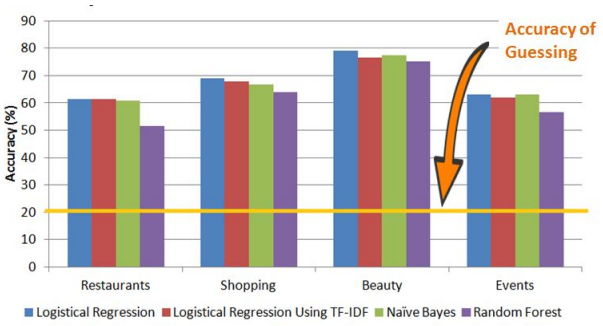

Fig. 1. Accuracy for each Classifier for each Category

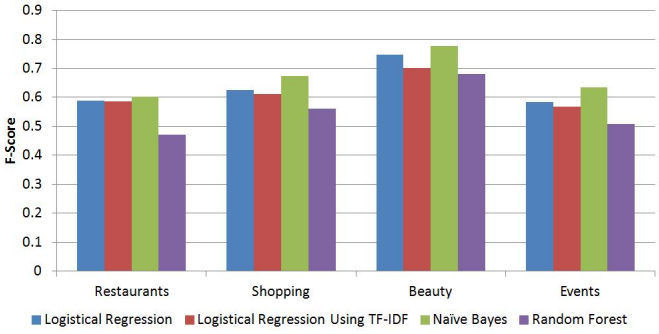

Fig. 2. F-score for each Classifier for each Category

Below are the AUCs for the logistic regression classifiers for each category. The remaining AUCs can be seen in VI.

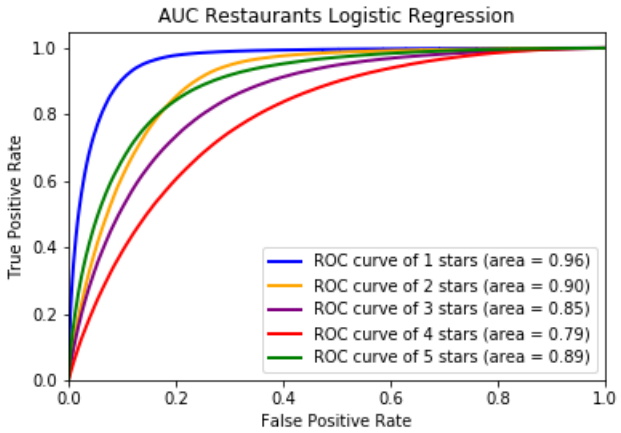

Fig. 3. AUC for Restaurants with Logistic Regression

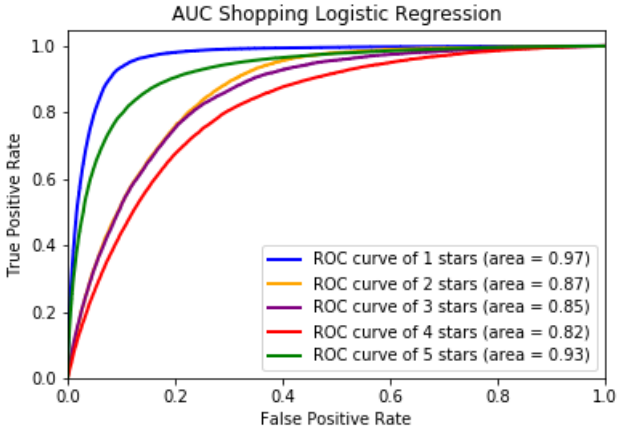

Fig. 4. AUC for Shopping with Logistic Regression

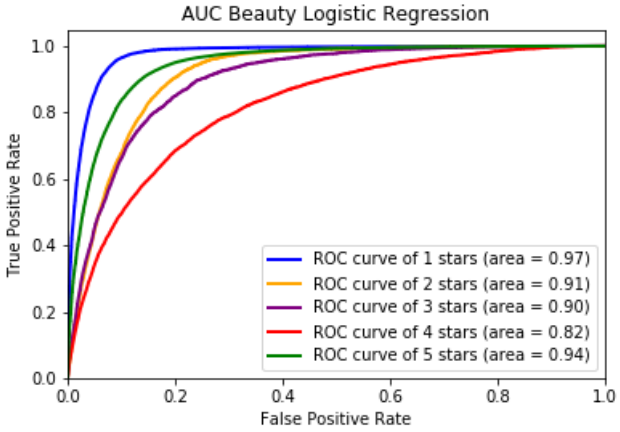

Fig. 5. AUC for Beauty and Spas with Logistic Regression

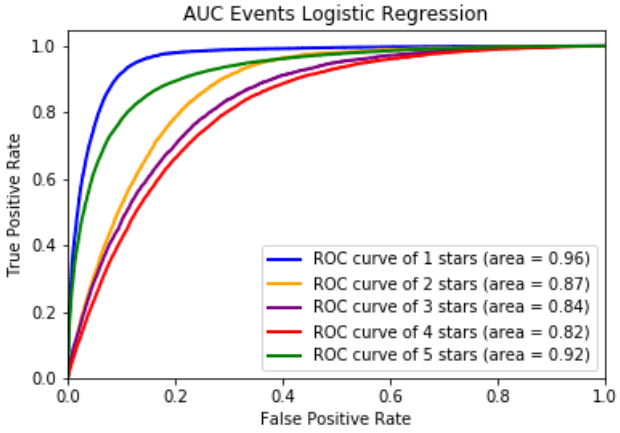

Fig. 6. AUC for Event Planning and Services with Logistic Regression

The confusion matrices for the logistic regression classifiers are shown below. The remaining confusion matrices can be seen in VII.

TABLE I
CONFUSION MATRIX FOR RESTAURANTS LOGISTIC REGRESSION

Actual \ Predicted	1	2	3	4	5
1	109,615	10,857	5,398	5,312	17,653
2	34,849	24,660	27,175	14,870	15,872
3	11,261	10,247	52,715	62,402	30,440
4	3,009	1,651	15,784	145,759	159,927
5	1,619	454	1,963	52,753	436,631

TABLE II
CONFUSION MATRIX FOR SHOPPING LOGISTIC REGRESSION

Actual \ Predicted	1	2	3	4	5
1	22,345	163	87	180	5,023
2	4,004	290	393	856	2,966
3	1,370	145	632	2,811	5,048
4	531	73	228	4,631	14,824
5	458	29	53	1,819	63,009

TABLE III
CONFUSION MATRIX FOR BEAUTY AND SPA LOGISTIC REGRESSION

Actual \ Predicted	1	2	3	4	5
1	13,726	451	96	82	2,318
2	2,866	720	278	221	1,354
3	887	382	424	779	1,847
4	263	103	203	1,797	7,824
5	247	51	63	852	64,027

TABLE IV
CONFUSION MATRIX FOR EVENT PLANNING AND SERVICES LOGISTIC REGRESSION

Actual \ Predicted	1	2	3	4	5
1	9,291	345	158	177	1,466
2	2,404	596	705	721	1,167
3	816	316	1,209	2,681	2,252
4	208	76	419	5,186	7,576
5	79	19	64	1,839	23,626

The other features which affect the star rating were determined, and can be seen graphically in the figure below. The location of a business (latitude and longitude) and the total number of reviews were seen to be the most influential.

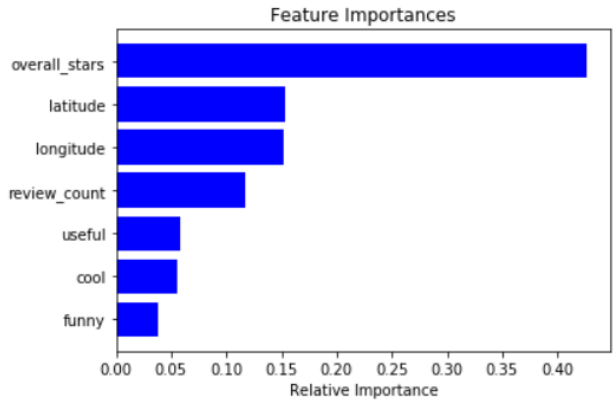

Fig. 7. F-score for each Classifier for each Category

The time to perform the data-cleaning process was recorded using different number of Spark nodes on the parallel cluster, in order to determine the effect of increasing the nodes.

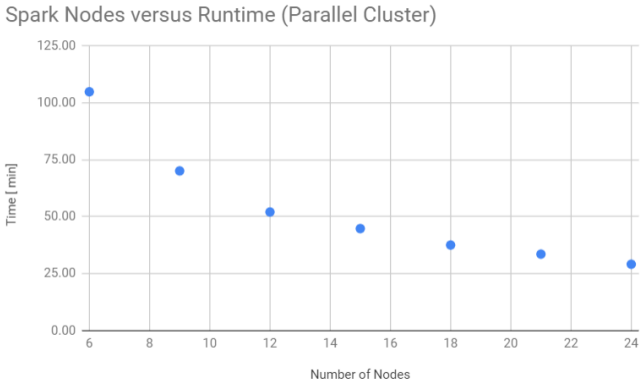

Fig. 8. Time to run the Cleaning Process for Different Node Amounts

IV. DISCUSSION

A. Limitations

In this study, there are a few limitations that need to be considered before drawing any final conclusions. Firstly, Yelp is used for many different types of companies. This is the reason that the data was split into different categories. However, many reviews had multiple categories for the business, which can change the context. For example, some hotel reviews were grouped into the restaurant category, as they have restaurants in them. The reviews would be focused mainly on the other aspects of the hotel, with a slight mention of the restaurant. These reviews will affect the corpus and make it more difficult to predict the star rating.

Reviews that are written in informal English can also present issues for the algorithms employed. The words may not be recognized and the review may be seen to not have any meaningful words. Therefore, it can difficult to accurately predict the star rating for the review.

Another limitation is that there are many different people writing the reviews, and there are different thoughts on what

each star rating implies. Some people will only give five stars if everything is absolutely perfect, whereas others will give it more freely. Additionally, the comments for four and five star reviews (or one and two stars) will tend to be fairly similar overall. Combined with the fact that there are many more five star reviews than four star reviews, this will lead to a bias in the prediction, as can be seen in the confusion matrices.

B. Analysis of Results

The restaurants category had by far the largest amount of data, while the other categories were relatively similar. The highest accuracy obtained for the restaurants category was 61.4%. The highest overall accuracy was 79.2% (logistic regression for beauty and spas). Combining the accuracies and weighing them appropriately gives an overall accuracy of 63.28%. This is very close to the accuracy found in the work done by Elkouri [8]. With more time, our models could likely be further optimized to provide an even higher accuracy.

As mentioned briefly in the limitations, the predictions were biased towards one and five star reviews, which leads to a decrease in accuracy. The predictions were often close to the correct star though (5 instead of 4, 1 instead of 2). This shows that the algorithms were able to correctly identify whether the sentiment was positive or negative accurately, but distinguishing between similar stars is difficult to do. When splitting the results into positive or negative sentiments (4-5 as positive, 1-3 as negative), our accuracy is 86.70%. This is 9% higher than the work of Jong [4].

As can be seen from the plot of nodes versus run time, there is an inverse relation between the number of nodes and the speed at which the cleaning task was completed. As more nodes were taken, the time was decreased less and less with each step.

C. Spark Use

Spark was used for all of the major parts of the project. First, the json files were uploaded to Spark and the data processing began. This involved reading in the files, combining the review and business files, filtering out non-English reviews, cleaning the text, and outputting the data to new json files to be used for machine learning. The filtering and cleaning was done with dataframe API and RDD operations.

The cleaned json files were then read into another notebook. Due to the nature of this project, there weren't many reviews filtered out, and the files were still to large to run on a single computer. Therefore, MLlib was used on Spark to perform the machine learning.

To tune the parameters of the machine learning models in order to increase the accuracy, there were 15-29 runs for each technique. It would not be possible to have done this without Spark, as there would not be enough time.

D. Possible Extensions [Future Work]

One addition to this work could be the standardization of star ratings based upon the comment given. If it is seen that a small percentage of users give 4 stars and leave comments very similar to common comments for 5 stars, the star rating could be adjusted.

This research could also provide the ability to automate star ratings for similar services. The user would enter a comment, and the star rating would be automatically assigned. This would avoid discrepancies between how individuals view associated star ratings.

During this study, it was determined that the location of a business and the total number of reviews a business had would also affect the star rating. However, this data was not used in order to assist with the prediction of the star rating. This is something that can be done in the future in order to obtain more accurate predictions.

V. Conclusion

We have used four different machine learning techniques to predict Yelp user's star ratings from the comments they provide. The different methods have all provided varying results, though the accuracies were fairly similar. Logistic regression had the highest across accuracy across all categories, as well as the best AUCs. We were able to obtain an overall accuracy of 63.28%, which is very close to the accuracy obtained in other projects. When classifying the text as positive or negative, we have an accuracy of 86.70%, which is quite a bit higher than what was done in other projects.

Additionally, we have determined other factors that would assist in the prediction of the star rating. The location of a business, as well as the total number of reviews a business had were seen to have a correlation with the star rating. This could be used in a future project in order to more accurately predict the star ratings.

Our findings could be useful in order to analyze how consistent the general public is with assigning a star rating to a specific type of experience. These findings could potentially be used in order to help standardize user's star ratings based on their comments to provide a more accurate rating for the business, making Yelp more reliable as to what experience the average customer will have at a particular business.

VI. Appendix A - AUCs

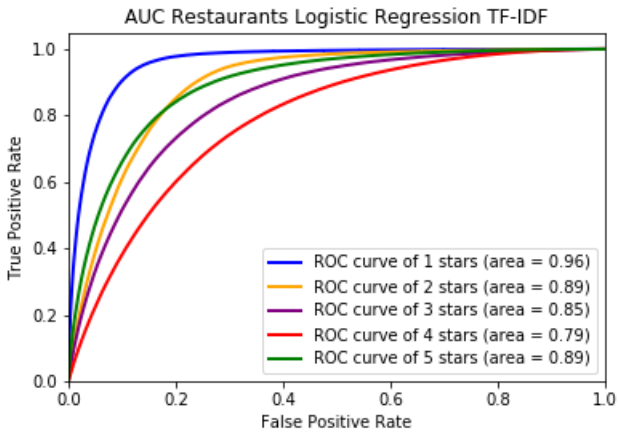

Fig. 9. AUC for Restaurants with Logistic Regression using TF-IDF

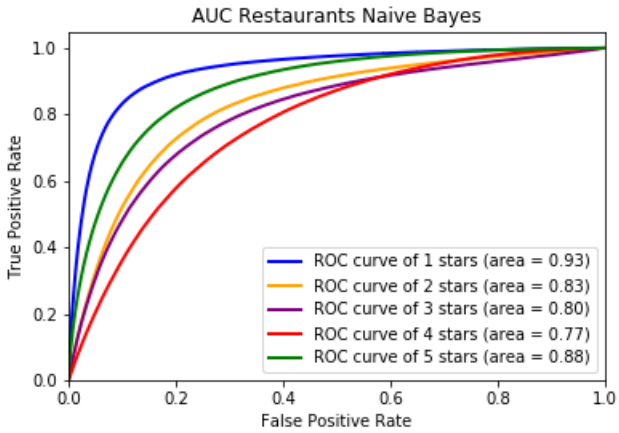

Fig. 10. AUC for Restaurants with Naive Bayes

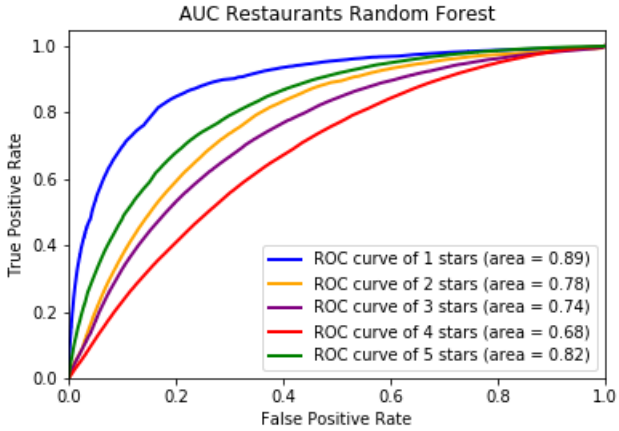

Fig. 11. AUC for Restaurants with Random Forest

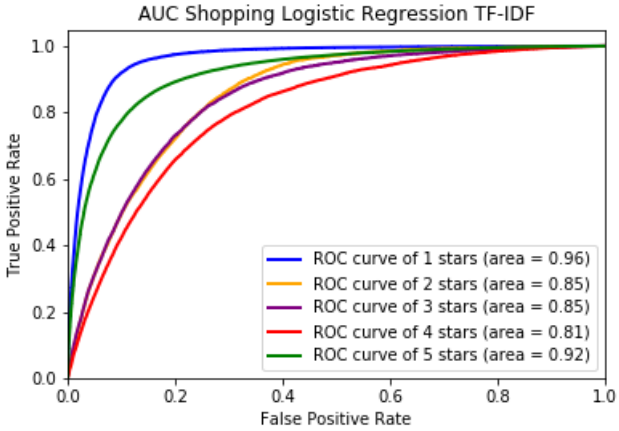

Fig. 12. AUC for Shopping with Logistic Regression using TF-IDF

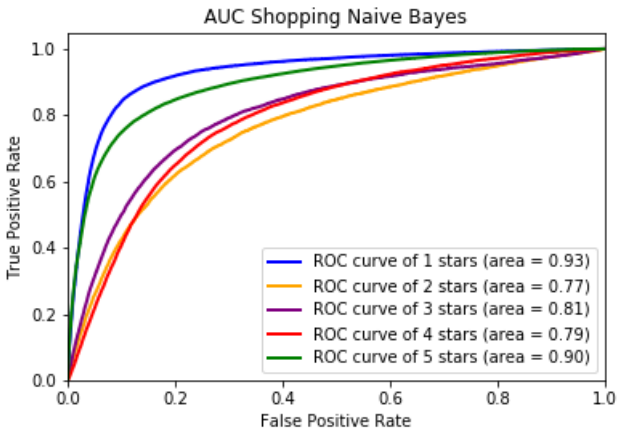

Fig. 13. AUC for Shopping with Naive Bayes

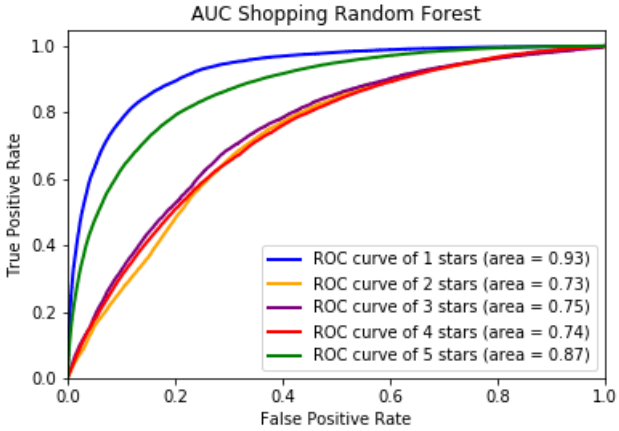

Fig. 14. AUC for Shopping with Random Forest

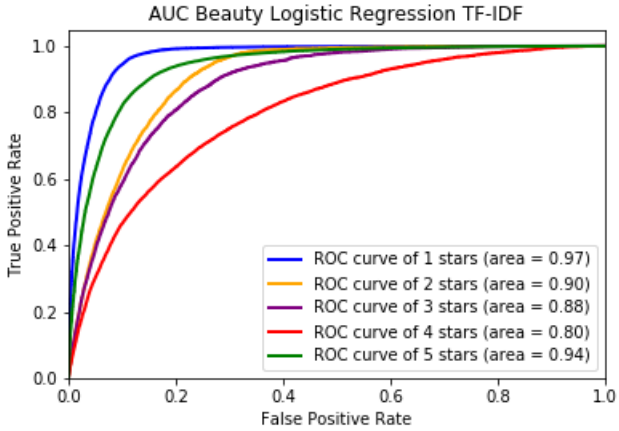

Fig. 15. AUC for Beauty and Spas with Logistic Regression using TF-IDF

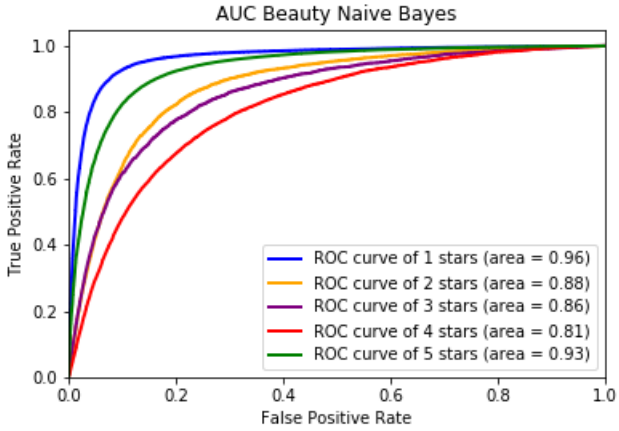

Fig. 16. AUC for Beauty and Spas with Naive Bayes

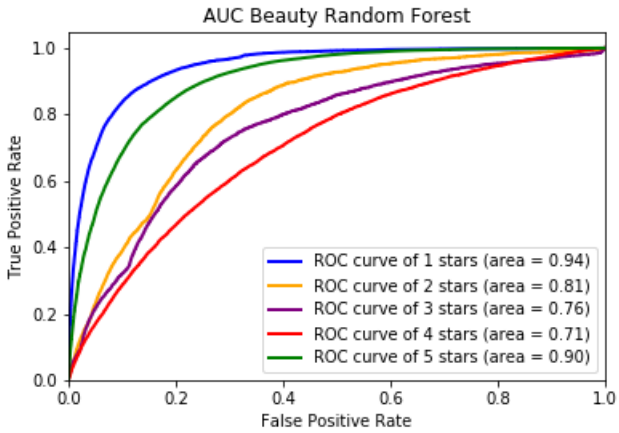

Fig. 17. AUC for Beauty and Spas with Random Forest

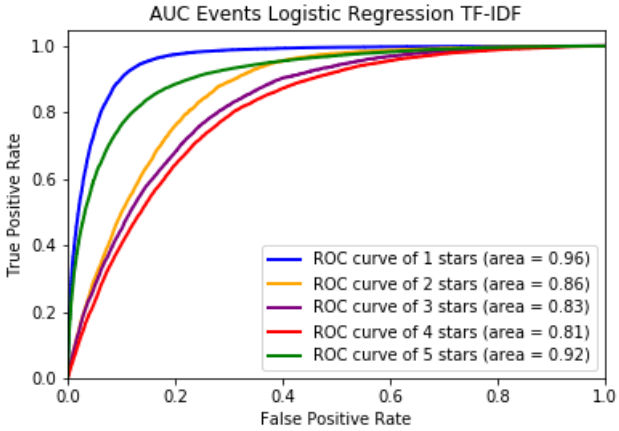

Fig. 18. AUC for Event Planning and Services with Logistic Regression using TF-IDF

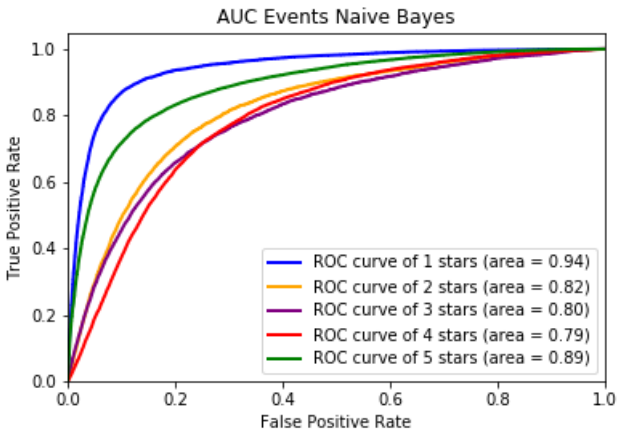

Fig. 19. AUC for Event Planning and Services with Naive Bayes

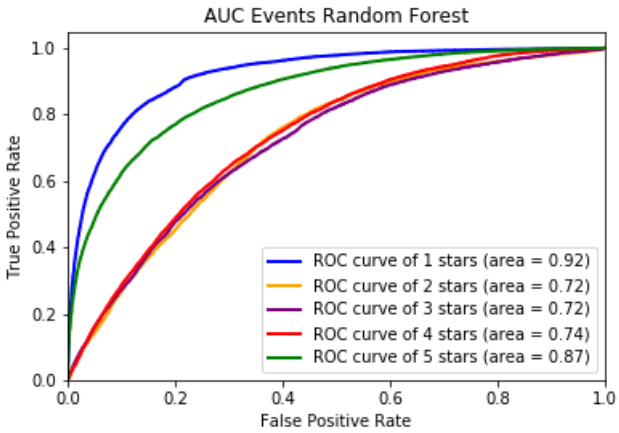

Fig. 20. AUC for Event Planning and Services with Random Forest

VII. Appendix B - Confusion Matrices

TABLE V
CONFUSION MATRIX FOR RESTAURANTS LOGISTIC REGRESSION TF-IDF

Actual \ Predicted	1	2	3	4	5
1	109,408	11,109	5,516	5,603	17,199
2	34,818	24,910	27,162	14,981	15,555
3	11,190	10,595	52,806	62,294	30,180
4	3,033	1,713	16,348	146,058	158,978
5	1,593	489	2,121	54,312	434,905

TABLE VI
CONFUSION MATRIX FOR RESTAURANTS NAIVE BAYES

Actual \ Predicted	1	2	3	4	5
1	107,263	28,550	7,665	2,438	2,919
2	33,461	40,338	32,634	7,261	3,732
3	15,426	23,077	68,773	46,646	13,143
4	7,991	8,527	37,212	151,976	120,424
5	8,461	3,815	8,410	78,722	394,012

TABLE VII
CONFUSION MATRIX FOR RESTAURANTS RANDOM FOREST

Actual \ Predicted	1	2	3	4	5
1	81,239	9,005	8,401	11,436	38,356
2	28,758	11,305	17,930	22,717	36,256
3	14,587	8,988	29,331	54,784	58,863
4	8,195	4,962	18,013	96,684	198,744
5	6,350	2,466	7,030	51,159	425,575

TABLE VIII
CONFUSION MATRIX FOR SHOPPING LOGISTIC REGRESSION TF-IDF

Actual \ Predicted	1	2	3	4	5
1	21,522	145	94	201	5,836
2	3,832	225	326	846	3,280
3	1,392	128	484	2,683	5,319
4	562	68	191	4,308	15,158
5	534	35	54	1,780	62,965

TABLE IX
CONFUSION MATRIX FOR SHOPPING NAIVE BAYES

Actual \ Predicted	1	2	3	4	5
1	21,738	3,362	1,299	515	884
2	2,874	2,025	2,139	923	548
3	1,136	1,072	3,263	3,656	879
4	951	656	2,265	11,046	5,369
5	2,723	609	909	11,222	49,905

TABLE X
CONFUSION MATRIX FOR SHOPPING RANDOM FOREST

Actual \ Predicted	1	2	3	4	5
1	19,818	171	101	248	7,460
2	3,383	127	189	534	4,276
3	1,580	93	333	1,297	6,703
4	1,177	85	286	2,040	16,699
5	1,811	48	113	1194	62,202

TABLE XI
CONFUSION MATRIX FOR BEAUTY AND SPAS LOGISTIC REGRESSION TF-IDF

Actual \ Predicted	1	2	3	4	5
1	12,087	153	25	58	4,350
2	2,722	303	56	196	2,162
3	886	205	97	527	2,604
4	281	67	48	905	8,889
5	252	41	15	378	64,554

TABLE XII
CONFUSION MATRIX FOR BEAUTY AND SPAS NAIVE BAYES

Actual \ Predicted	1	2	3	4	5
1	13,616	1,915	473	224	445
2	2,196	1,862	713	377	291
3	665	969	1019	1,145	521
4	337	403	733	4,298	4,419
5	654	400	319	5,758	58,109

TABLE XIII
CONFUSION MATRIX FOR BEAUTY AND SPAS RANDOM FOREST

Actual \ Predicted	1	2	3	4	5
1	12,036	214	44	87	4,292
2	2,612	190	63	107	2,467
3	1,047	99	80	151	2,942
4	593	73	58	250	9,216
5	824	46	29	196	64,145

TABLE XIV
CONFUSION MATRIX FOR EVENT PLANNING AND SERVICES LOGISTIC REGRESSION TF-IDF

Actual \ Predicted	1	2	3	4	5
1	8,999	289	142	210	1,797
2	2,337	496	653	755	1,352
3	847	272	1,062	2,627	2,466
4	214	100	377	5,028	7,746
5	107	16	63	1,824	23,617

TABLE XV
CONFUSION MATRIX FOR EVENT PLANNING AND SERVICES NAIVE BAYES

Actual \ Predicted	1	2	3	4	5
1	8,642	1,764	549	233	249
2	1,580	1,837	1,504	430	242
3	607	1,026	2,671	2,412	558
4	266	415	1,646	8,070	3,068
5	319	164	368	6,083	18,693

TABLE XVI
CONFUSION MATRIX FOR EVENT PLANNING AND SERVICES RANDOM FOREST

Actual \ Predicted	1	2	3	4	5
1	8,229	278	242	451	2,237
2	2,151	266	392	910	1,874
3	1,132	207	712	2,100	3,123
4	574	156	567	3,855	8,313
5	508	64	205	2,074	22,776

REFERENCES

[1] M. Anderson and J. Magruder, "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database," *The Economic Journal*, vol. 122, no. 563, pp. 957–989, 2012.

[2] Yelp, "Yelp open dataset: An all purpose dataset for learning," 2019. [Online]. Available: https://www.yelp.com/factsheet

[3] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content," in *WebDB*, 2009.

[4] J. Jong, "Predicting rating with sentiment analysis," 2011.

[5] Y. Xu, X. Wu, and Q. Wang, "Sentiment analysis of yelps ratings based on text reviews," 2014.

[6] N. Asghar, "Yelp dataset challenge: Review rating prediction," *CoRR*, vol. abs/1605.05362, 2016. [Online]. Available: http://arxiv.org/abs/1605.05362

[7] K. Carbon and A. Stanford, "Applications of machine learning to predict yelp ratings," 2014.

[8] A. Elkouri, "Predicting the sentiment polarity and rating of yelp reviews," *CoRR*, vol. abs/1512.06303, 2015. [Online]. Available: http://arxiv.org/abs/1512.06303

[9] M. Fan and M. Khademi, "Predicting a business star in yelp from its reviews text alone," *CoRR*, vol. abs/1401.0864, 2014. [Online]. Available: http://arxiv.org/abs/1401.0864