Relation Between Elite and Non-Elite User Sentiments and Business Rating Classification Using Yelp Data

Shikhar Srivastava
x18106960

***Abstract*—Hotel industry is a huge industry and unlike old days, this industry nowadays depend a lot on customer feedback. This setup of giving and receiving feedback on services provided in hotel industry not only benefit businesses to grow and improve their services, but also helps customers to select a hotel or restaurant of their choice based on feedback which a respective business has received in the past. There are many online platforms through which customers can check the reviews of a business. A group of people known as influencers are considered very important on these platforms and their reviews might impact customers decision to choose a hotel or restaurant. We are using business, review and user dataset provided by Yelp on their website. In these datasets, so called influencers are termed as elite users. Our data mining project is based on analyzing the impact of elite user reviews on other users for different businesses in the state of Ohio, US. We are further building a model which can predict the overall star rating of businesses based on sentiments calculated from all reviews.**

***Index Terms*—Yelp User Reviews, Impact of Elite Users, Business Rating Prediction, KNN, Linear Regression, Sentiment Analysis.**

I. Introduction

As the number of restaurants and hotels are increasing rapidly nowadays, it sometimes become very difficult to choose from a wide variety of options available. To make this task of selecting a restaurant or a hotel less painful and time taking, today one can easily check reviews of multiple businesses on few online platforms and select one of their choice. Its not just the customers who benefit from this review governed approach but also the businesses. Businesses can make use of these reviews posted on online platforms and improve their services to achieve higher customer satisfaction levels which in long run will result in higher profits. With online reviews starting to become more and more important every day for businesses and customers, a group of people called influencers have also become important because the reviews they post on online platforms are considered widely to influence choices made by customers in selecting a place to visit. Yelp is a popularly known online platform where users can look at the reviews of businesses and make their choice. The group of people who are ideally considered to be the influencers are termed as Elite users in Yelp.

For our project, we are using publicly available Yelp dataset which is present on their website. We first carry out a detailed exploratory analysis of data and then our project tries to analyze the relationship between the reviews given by elite users and the reviews given by normal users. This analysis is based on the sentiments of reviews posted by elite users and normal users. The relationship between them is checked by building a linear regression model. The sentiments are calculated by using nrc lexicons which categorizes words in 10 different emotional categories like anger, anticipation etc. So our first research question is ***"Do reviews given by elite users have any impact on reviews given by normal users?"***. This project not only just analyzes the relationship between elite and normal user reviews but also tries to build classification models that are able to predict star rating of a business based on the sentiments calculated from reviews of all users. For achieving this we are building different models like K-Nearest Neighbor, SVM and C5.0. This leads to our second research question ***"Which model among KNN, SVM and C5.0 can classify star rating of businesses and achieve better results?"***. We achieved 61% accuracy in predicting the star rating for businesses using KNN and 61%, 57% using SVM, C5.0 respectively on date when models were built. Due to huge class imbalance between the classification levels, we also made use of under-sampling method and implemented it before feeding our data in models.

Our project document is divided in 3 parts i.e. Data Mining Methodology, Evaluation & Results and Conclusion. Data Mining Methodology is further divided in 3 sub-sections, Data Acquisition and Pre-processing, Data Exploration and Models. In data acquisition and pre-processing we are explaining the source of data and all the pre-processing tasks done on data. In data exploration, few graphs are built to perform an exploratory analysis on data. The models sub-section covers all the steps we followed for building our models. Evaluation & Results provides a detailed explanation and comparison of all the results we got while building different models. In the last section, we are concluding our project.

II. Related Work

In today's fast paced world, people are faced with information overload. More and more businesses have presence on internet. As a result, it becomes difficult for users to choose the best option. An easy way to assist users in this selection process is provision of easy to understand indicators. Many popular sites like Yelp, TripAdvisor make use of user ratings and reviews for this purpose. Online reviews about business can have impact on the users. [1] So it is not a surprise that there is a type of marketing that makes use of Social Media Influencers. Businesses make use of these influencers and their large followers for advertisement. For the setting of Yelp, Elite users can be considered to be minor celebrities or influencers. There have been studies like - [2] - [3] - [4] around this as well. To understand the influence of the customer opinions/reviews on the decision making, while product purchase is made various studies have been conducted. One of such paper discussed how the customer review impact sale of the product. This research found a positive relationship between the average rating and sales. [10] Other researchers [11] found that variance or the standard deviation of the customer rating is more useful than the average rating to predict the sales growth. It is quite intuitive that the sales of the products decrease as a result of the higher number of negative reviews.

Along with sentiments in the review, importance of length of the review, the quality of language and selection of the words, has been studied in - [12]. This study showed that stylistic element may have more impact on customers than the length of the review comment.

The studies around online review analysis mainly make use of natural language processing. Text mining and especially Natural language processing has been studied and is being perfected since last decade. A comprehensive insight in the state of the art natural language processing techniques using deep learning has been done by Young and others in their paper - [8]. Text mining is useful for finding the patterns from the dataset. It can be effectively performed on unstructured or semi-structured datasets. Text mining is fundamentally the automatic process of analysing the various textual resource. As discussed by the author [9], it is divided into three major component Information feeders which includes information about the resources, Intelligent Tagging component perform statistical or semantic or structural tagging, and the final component - Business Intelligence suite which is responsible for combining information from the various sources and simulation analysis. In most of the text mining approach, the document can be represented by a set of words. Two main categories are task-based - the formal Frameworks and Algorithm-Based Techniques. In task-oriented approach, the problem is divided into a subset of the task. This subset is divided into three main class such as preparatory processing, general purpose NLP tasks, and problem dependent tasks. Formal Frameworks and Algorithm-Based Techniques includes Text categorization and Probabilistic models for Information Extraction. Further, text categorization can be of two types by defining the rules manually or with the help of machine learning which builds automatic text classifier by learning from the pre-classified set of training dataset. [9]

With huge amount of data, text analysis becomes a challenging task. Because of the various parameter of the Big data, it Is important that computational storage and data representation work together. This improves the search and mining in the large text collection. Data velocity and veracity can affect the results. Along with this, data quality and data cleaning are also some of the major challenges in context to text mining [9]

Khoo and others studied how these online reviews given on the site TripAdvisor affect the users decisions. [4] This study used top ten frequent word counts for positive and negative emotions using Wmatrix. Then human raters were used to rate hotels based on their perception of these reviews. A similar study, but without involvement of human raters - [5] tried to predict star rating from the review text. This research used word frequency to quantify the review text. Four regression models, namely, linear regression, support vector regression - with and without normalized features and decision tree regression were used for prediction. Total review count used was around 36000. This study used highest occurring words in the dataset, which does not necessarily represent the sentiment of the reviews in a precise way.

In a study - [6], Li and others used an approach of sentiment analysis which yields more categories of sentiments than the simple binary - positive and negative. This is especially important as there is a possibility that two or more reviews can have positive (or negative) review but of different intensity. The dataset used in this study was scrapped from site epinions.com. Content based model was utilised for sentiment analysis in this study. However, there are more refined libraries to perform this task using R. This study utilised Linear Regression and Support Vector Machine to predict review rating. One of the important rating that users consider to be important is the rating of business which was not considered or predicted in this study.

Along with other machine learning methods, neural networks were used by Tang and others in a study - [7], for user rating prediction using sentiment analysis of the user reviews. This study used two datasets, Yelp dataset for restaurant reviews and Rotten Tomatoes dataset for movie reviews. Deep learning was used to perform sentiment analysis. An interesting approach was followed in this paper to consider user specific meaning of the word. This study proposed and used a model called user-word composition vector model [7]. The approach used to give user specific meaning to words was built using matrix multiplication. Each user was assigned a matrix which was then used for multiplication and ultimately to modify meaning of certain word. The most novel and intriguing aspect of this study was the inclusion of user data along with review sentiment in the rating prediction.

III. Data Mining Methodology

This section provides a detailed explanation of data used in this project along with steps followed to prepare data. We are further performing an exploratory analysis on the dataset and explaining models which we built to answer our second research question. Data mining methodology followed for this project is KDD - Knowledge Discovery and Data Mining.

A. Data Acquisition and Preprocessing

Dataset used in this project has been downloaded from Yelp website which is publicly provided by Yelp, so there are no ethical issues associated with our dataset. Although on their website Yelp provides a wide range of datasets in the form of JSON files which includes Business.json, Review.json, User.json, Checkin.json, Tip.json and Photo.json. But, for this project we are only using Business.json, Review.json and User.json. Zipped file of size 3.6 GB containing all the jsons were downloaded after which we used only the relevant files as mentioned above.

After downloading Business.json, Review.json and User.json, these files were then converted to csv files (R. Code1_JSON_to_CSV.R). Though there are many columns present in json files, we are only selecting columns relevant to our project using the same code. Our dataset was huge to handle it in R so, for answering our research questions we filtered our dataset with businesses present in the state of OHIO only. The converted csv files were then binded together to form a single file having all users, businesses and reviews in the state of OHIO. We then mutated an attribute named elite_on_review_date in our dataset which distinguishes elite users from normal users, for building a linear regression model to answer our first research question (Code2_Script.R). Just to make sure that the preprocessing task didn't altered our data in terms of volume, we noted number of observations before preprocessing and then after preprocessing which came out to be exactly same i.e. 172477 observations.

For building our linear regression model to find relationship intensity between reviews given by elite users and reviews given by normal users, we are first calculating the sentiments of reviews given by elite users and sentiments of reviews given by normal users using nrc lexicons present in Code3_Sentiment_Elite.R and Code4_Sentiment_nonelite.R respectively. We also calculated sentiments for building our KNN, SVM and C5.0 models. For using the sentiments in our models, we are first calculating sentiments for all reviews and for every business and then taking mean of sentiment scores grouped by respective businesses (Code5_adm_yelp.R). But, before calculating any sentiments from reviews to answer our first and second research questions, we are first cleaning all the reviews which involves removal of whitespace, removal of punctuation marks, removal of stopwords etc. (Code2_Script.R). Especially for building the models to classify star rating, apart from just cleaning our reviews, we are also performing under-sampling technique to achieve balanced classes (Code5_adm_yelp.R). Last but not the least, we are also creating a separate column to represent business star ratings which is basically a categorization of already existing business star ratings and is categorized by 3 levels i.e. Low, Medium & High (Code5_adm_yelp.R).

B. Data Exploration

An exploratory analysis of our data was carried out to understand it and get an overview. Figure 1 and 2 show the internal structure and summary of our data respectively.

A bar chart was built to get an idea of top 10 cities in the state of Ohio having most number of businesses as shown in Figure 3. It was noticed that Cleveland was the only city in Ohio where most of the businesses are concentrated and number of unique businesses operating in Cleveland was way to high as compared to other cities of Ohio.

Another chart was build to get an idea about how many reviews have been given by elite and non-elite users. This was represented again by a simple bar chart as shown in Figure 4 and shows total number of reviews given by elite users to be 36910 and total number of reviews given by non elite users to be 135567.

```
'data.frame':   172477 obs. of  13 variables:
 $ user_id           : chr  "--_EpULz-cjQit4npXy1ng" "--_EpULz-cjQit4npXy1ng" "--_EpULz-cjQit4npXy1ng" "--_EpULz-cjQit4npXy1ng" ...
 $ business_id       : chr  "yXztaR7Flo2qk_70XuBxZw" "5YyLWps6xL372Lj2hRlPQw" "qO7dpBAVxbrR4GbtR2uCOw" "du9EiKx28BVs-2hUAQ3UVg" ...
 $ review_id         : chr  "mCUJ3fwMFu6OLlE7-9iGDA" "DNeZNOvqZvIERrQYmRC_Dg" "A2kGwP2a6SZwnctJ7KL6zg" "VWObSapALHanLPned5yn8A" ...
 $ review_stars      : int  5 5 1 5 5 3 5 4 2 5 ...
 $ date              : chr  "2013-10-16 06:34:41" "2013-10-16 06:30:45" "2015-09-07 22:35:46" "2014-08-19 23:12:20" ...
 $ text              : chr  "Moved away from North Olmsted and I am missing my favorite go to restaurant.  The wait staff was always good.  "| __trun
cated__ "I love the wings they are just as good as my hometown one in PA.  I had the best anywhere there and missed them"| __truncated__ "My husband an
d I went into check the new restaurant in Avon on a very hot September day.  The place wasn't busy"| __truncated__ "excellent service and food!!  serve
r was great.  food was phenomenal!" ...
 $ business_name     : chr  "Chili's" "Quaker Steak & Lube" "Heck's Café" "Cafe Melissa" ...
 $ city              : chr  "North Olmsted" "Sheffield Village" "Avon" "Avon Lake" ...
 $ business_stars    : num  3 3 3.5 4 3.5 4.5 4.5 3.5 3.5 4 ...
 $ state             : chr  "OH" "OH" "OH" "OH" ...
 $ user_name         : chr  "Pat" "Pat" "Pat" "Pat" ...
 $ elite             : chr  "" "" "" "" ...
 $ elite_on_review_date: chr  NA NA NA NA ...
```

Fig. 1. Structure of Dataset

```
   user_id          business_id         review_id          review_stars        date               text            business_name     
 Length:172477      Length:172477      Length:172477      Min.   :1.000   Length:172477      Length:172477      Length:172477     
 Class :character   Class :character   Class :character   1st Qu.:3.000   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Median :4.000   Mode  :character   Mode  :character   Mode  :character  
                                                          Mean   :3.665                                                           
                                                          3rd Qu.:5.000                                                           
                                                          Max.   :5.000                                                           
     city           business_stars     state             user_name             elite           elite_on_review_date
 Length:172477      Min.   :1.000   Length:172477      Length:172477      Length:172477      Length:172477     
 Class :character   1st Qu.:3.500   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Median :4.000   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                    Mean   :3.661                                                                              
                    3rd Qu.:4.000                                                                              
                    Max.   :5.000                                                                              
```

Fig. 2. Summary of Dataset

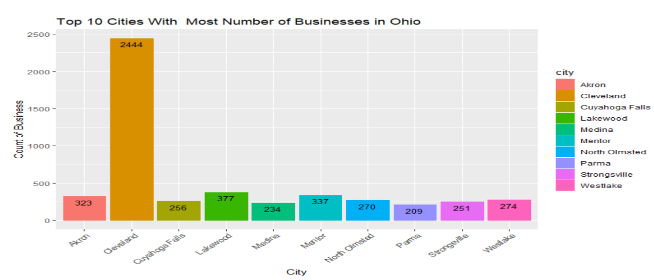

Fig. 3. Top 10 Cities With Most Number of Businesses in Ohio

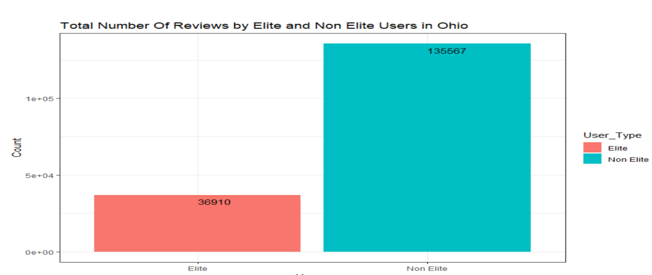

Fig. 4. Total Number Of Reviews by Elite and Non Elite Users in Ohio

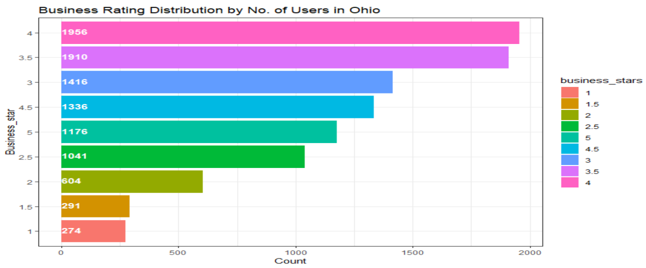

Fig. 5. Business Rating Distribution by No. of Users in Ohio

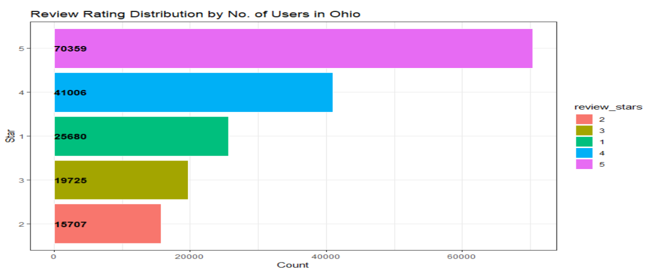

Fig. 6. Review Rating Distribution by No. of Users in Ohio

As the star ratings are an important aspect in this project, so we also built a chart showing number of users by star rating. In our dataset, we have two types of ratings, one is the overall business rating and other shows rating given by each user on the basis of their review. In Figure 5 and 6 we tried to identify number of users categorized by rating they gave for each rating type i.e. overall business rating and review rating. It was noticed that for business rating most of the businesses achieved high overall rating of 4 with 1956 businesses falling under this category while majority of businesses received an overall rating of 3, 3.5, 4, 4.5. For ratings given by users on basis of their reviews, it can be seen that most of the users gave rating of 4 or 5. Interestingly, number of users who gave rating 1 is higher than number of users who gave rating 2 and 3.

Because our project deals with calculating sentiments from reviews, we thought that building a word cloud for top 100 most frequently used words in reviews would be really insightful. Figure 7 shows a wordcloud which we built in R. As we can see in the image, FOOD, GOOD, GREAT, PLACE are some of the most frequently used words in reviews given by users in Ohio.

C. Models

1) First Research Question: A linear regression model was built first to answer one of our research questions which tries to find a relationship between reviews given by elite users and reviews given by non-elite users. As stated in pre-processing section, we mutated a column to distinguish elite user reviews from non-elite user reviews. This mutated column was then used to create 2 subsets of data for elite and non-elite. For each business, elite user sentiments and non-elite user sentiments were calculated using nrc lexicons. Here sentiments

Fig. 7. 100 Most Frequently Used Words

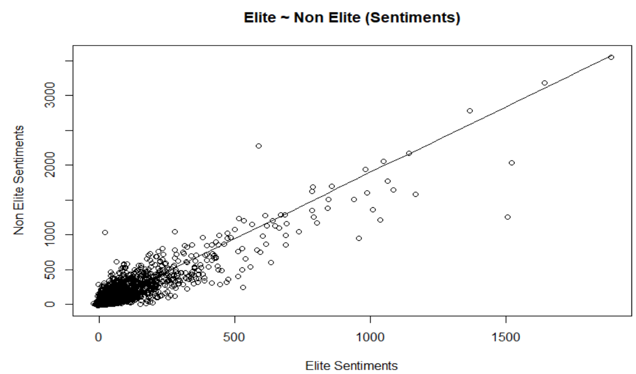

Fig. 8. Scatter Plot

are nothing but a difference of positive and negative category present in nrc package. Figure 8 shows a scatter plot built to get an overview of relationship between elite user review sentiments and non elite user review sentiments. A very high correlation of 90.7% was noticed between elite user review sentiments and non elite user review sentiments. To further check the impact of elite user reviews on non-elite users, we then built a linear regression model.

2) Second Research Question: To answer our second research question i.e. to be able to build a model which can classify businesses correctly into 3 different categories (as explained in pre-processing section), we are building 3 different models and then comparing their performance based on sensitivity and specificity. As stated in previous sections, we are building KNN, SVM and C5.0 classification models to classify businesses into Low, Medium and High category.

The entire code used to answer our second research question is present in Code5_adm_yelp.R. Before feeding our data to any of the above mentioned models we are first calculating sentiments of every user review followed by taking an average of all sentiments and grouping it by each business, next we are categorizing already existing business ratings into Low, Medium and High category. Then under sampling is done on our data to maintain the class balance. After all this, our data is

```
lm(formula = Nonelite_Sent ~ Elite_Sent, data = elite_nonelite_sent)

Residuals:
     Min       1Q   Median       3Q      Max
-1171.08   -24.28    -9.89    13.91  1322.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.891984   0.997665   14.93   <2e-16 ***
Elite_Sent   1.599063   0.009526  167.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.97 on 6069 degrees of freedom
Multiple R-squared:  0.8228,	Adjusted R-squared:  0.8228
F-statistic: 2.818e+04 on 1 and 6069 DF,  p-value: < 2.2e-16
```

Fig. 9. Linear Regression Model Output

then fed to KNN, SVM and C5.0 models. In case of building the KNN classification model, we carried out two additional tasks, first being finding the best K value for our model ranging from 1 to 100. We did this by using a for loop and recording accuracy for each value of K from 1 to 100. Second, to avoid bias and effects of variance in the test data we performed K-Fold Cross validation technique and took mean accuracy of all folds.

IV. Evaluation and Results

In this section the output of selected models in the methodology are analysed and compared.

A. Elite User Sentiments Influence Regular Users Sentiments

We built a linear regression model to measure the relationship of elite user sentiments with non-elite users sentiments which answers our first research question. Figure 9 shows our linear regression model output we got in R. The equation we obtained for our regression model is Y = 1.599063X + 14.891984, where Y represents our non-elite user sentiments and X represents our elite user sentiments. Looking at the R-squared value we can state that 82.2% total variation in the non-elite user sentiments (Y) is explained or accounted for by the variation in the elite user sentiments (X). The equation means that for every unit increase in elite user sentiments, non-elite user sentiments increases by 1.599063. This clearly shows that there is considerable amount of impact of elite user sentiments on non-elite user sentiments which correctly answers our first research question.

B. Business Rating Based in Reviews Sentiment Analysis

The first proposed method in this study for classification of business rating was the non parametric model and one of the simplest machine learning algorithms, KNN. We iterated he number of neighbours (K) from 1 to 200 and the best accuracy was achieved with K = 33. To avoid bias and variance from the training model we used a K-fold cross validation approach. The overall accuracy achieved with this technique was 60% on average of 20 windows. From the data set of testing the Low was correctly classified in 67% of the observations, the 49% of the cases from Medium were correctly classified while the percentage from correctly classification of High was 62%. The error rate of this model is 40% which leads us to try other techniques.

TABLE I
K-Nearest Neighbors Results

Model	KNN		
Accuracy	60%		
Kappa	0.40		
	Low	***Medium***	***High***
Sensitivity/Recall	0.67	0.49	0.62
Specificity	0.85	0.73	0.81
Precision	0.71	0.49	0.58
F1-score	0.69	0.49	0.60
Balanced Accuracy	0.76	0.61	0.72
Confusion Matrix			
	Reference		
Prediction	***Low***	***Medium***	***High***
Low	399	118	41
Medium	145	281	145
High	50	168	309

TABLE II
Support Vector Machine

Model	SVM		
Accuracy	61%		
Kappa	0.41		
	Low	***Medium***	***High***
Sensitivity/Recall	0.66	0.54	0.61
Specificity	0.86	0.72	0.84
Precision	0.74	0.41	0.68
F1-score	0.69	0.49	0.60
Balanced Accuracy	0.76	0.63	0.72
Confusion Matrix			
	Reference		
Prediction	***Low***	***Medium***	***High***
Low	416	87	55
Medium	161	234	176
High	54	116	357

The choice of SVM as a second model for business rating classification was given because of the hability of this method in dividing the space in hyperplanes separating each Class of the data, performing also non-linear classifications. The overall accuracy achieved using SVM was over 61% with the randomly selected observations. From the data set of testing the Low was correctly classified in 66% of the observations, the 54% of the cases from Medium were correctly classified while the percentage from correctly classification of High was 61%. The error rate of this model is 40%. This approach increased in 5% the correct classification of observations of the Medium.

The third model used for the business rating classification in this study was Decision Tree, also known as C50. Decision trees also as KNN are one of the standard and simplest algorithms for classification methods. The overall accuracy achieved using this method was 55% using randomly selected observations. The rate of cases correctly classified as Low was 62% from the training data set, the 46% of the observations

TABLE III
C5.0 Decision Trees and Rule-Based Models

Model	C50		
Accuracy	55%		
Kappa	0.33		
	Low	***Medium***	***High***
Sensitivity/Recall	0.62	0.46	0.55
Specificity	0.85	0.70	0.79
Precision	0.72	0.39	0.54
F1-score	0.67	0.43	0.55
Balanced Accuracy	0.73	0.58	0.67
Confusion Matrix			
	Reference		
Prediction	***Low***	***Medium***	***High***
Low	404	87	67
Medium	179	225	167
High	68	171	288

from High were correctly classified and the percentage of right classifications of High was 55%. This model has the higher error rate 45% among the other classifiers.

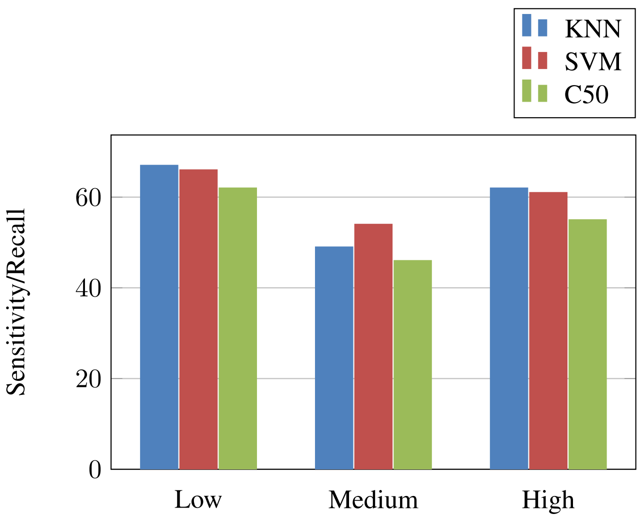

Fig. 10. Sensitivity of Models

In the above graphic it is possible to compare and analyse the sensitivity (recall) of the three selected models. The KNN and SVM presented very similar rate of correctly classifications of Low and High. It is also possible to observe that the most effective method for classification of elements from Medium is SVM.

Another very important metric for evaluation of machine learning models is presented in the Specificity graphic. As we can see the three models are quite similar at correctly classifying elements as not from Low and not from Medium. SVM is slightly better at classifying elements as not belonging to High.

V. Conclusion

Customer feedback has nowadays become very important for every business and businesses actually rely a lot on reviews given by their customers. The customers can now give an

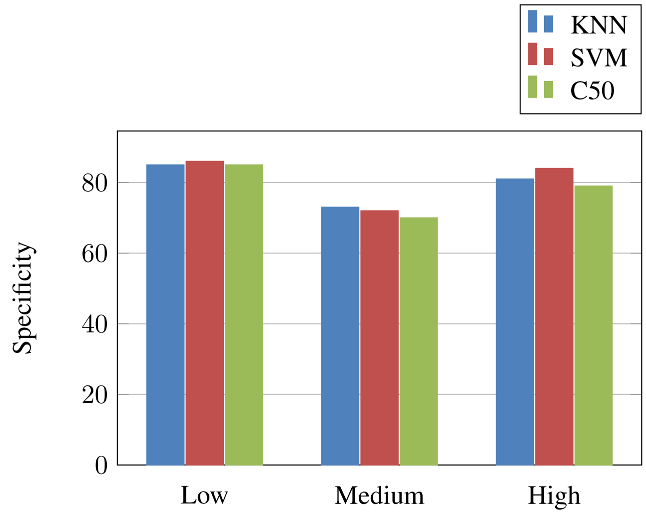

Fig. 11. Specificity of Models

un-biased feedback on several online platforms platforms. A group of people called influencers are considered to be an important group of people who might have an impact on other users. Reviews given these people can impact businesses in a positive or a negative way. These platforms also enable users to give star rating to businesses based on their experience which is also very important because it not only helps other users to choose best restaurant/hotel by looking at the star rating, but also helps businesses to improve their services.

Our project used data provided by Yelp to answer our 2 research questions. We built a linear regression model to check the relationship between sentiments of elite user reviews and sentiments of non-elite user reviews. We found that the R-Squared value for our linear equation (where non-elite user sentiments is our dependent variable and elite user sentiments are independent variable) is 0.8228 which means that 82.2% total variation in the non-elite user sentiments is explained or accounted for by the variation in the elite user sentiments and p-value ¡ 2.2e-16 which shows that there is a solid relationship between elite user sentiments and non-elite user sentiments and elite user sentiments impact non-elite user sentiments.

To answer our second research question, we built 3 different models i.e. KNN, SVM and C5.0 for classification of businesses into Low, Medium and High classes (derived from overall star rating of businesses). We found that among the 3 models built, SVM performed the best with an accuracy of 61% with Kappa value of 0.41. Lowest performance was recorded for C5.0 model (SVM (61%) > KNN (60%) > C50 (55%)).

For future work, this classification can be done using more sophisticated techniques like Neural Networks. For our project we only used Yelp data for the state of Ohio, so in any future analysis the entire dataset can be taken. For our SVM model used for second research question we used Laplacian kernel, future studies can also try other kernel types. Future studies can also build an ensemble model to get achieve performance metrics. Probably, a boosting algorithm can be used like XGBoost to improve accuracy for our models built to answer second research question.

REFERENCES

[1] J. Luan, W. Shan, Y. Wang, and J. Xiao, How easy-to-process information influences consumers over time: Online review vs. brand popularity, Computers in Human Behavior, vol. 97, pp. 193201, 2019.

[2] C. Abidin, Visibility labour: Engaging with Influencers fashion brands and #OOTD advertorial campaigns on Instagram, Media International Australia, vol. 161, no. 1, pp. 86100, Nov. 2016.

[3] A. Cunningham, An investigation into consumer and influencer perceptions of using social media influencers at events as a way to promote fashion and beauty brands., Thesis, 2018.

[4] F. S. Khoo, P. L. Teh, and P. B. Ooi, CONSISTENCY OF ONLINE CONSUMERS PERCEPTIONS OF POSTED COMMENTS: AN ANALYSIS OF TRIPADVISOR REVIEWS, no. 2, p. 20, 2017.

[5] M. Fan and M. Khademi, Predicting a Business Star in Yelp from Its Reviews Text Alone, p. 5, Jan. 2014.

[6] F. Li, N. Nan Liu, H. Jin, K. Zhao, Q. Yang, and X. Zhu, Incorporating Reviewer and Product Information for Review Rating Prediction., presented at the IJCAI International Joint Conference on Artificial Intelligence, 2011, pp. 18201825.

[7] D. Tang, B. Qin, T. Liu, and Y. Yang, User Modeling with Neural Network for Review Rating Prediction, in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[8] T. Young, D. Hazarika, S. Poria, and E. Cambria, Recent Trends in Deep Learning Based Natural Language Processing [Review Article], IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 5575, Aug. 2018.

[9] N. Archak, A. Ghose, and P. G. Ipeirotis, Working Paper CeDER-07-05, New York University Deriving the Pricing Power of Product Features by Mining Consumer Reviews, May. 2011.

[10] N. Hu, P. Pavlou, and J. Zhang, Overcoming the J-Shaped Distribution of Product Reviews, Communications of the ACM, vol. 52, pp. 144147, Oct. 2009.

[11] Y. Pan and J. Q. Zhang, Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews, Journal of Retailing, vol. 87, no. 4, pp. 598612, Dec. 2011.

[12] C. Forman, A. Ghose, and B. Wiesenfeld, Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets, Information Systems Research, vol. 19, pp. 291313, 2008.