RELIABILITY DIMENSION IN REVIEW PLATFORM: YELP

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

BACHELOR OF TECHNOLOGY

in

Mathematics and Computing

by

Akash Mahalik

(Roll No. 150123004)

Tarun Genwa

(Roll No. 150123043)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

April 2019

CERTIFICATE

This is to certify that the work contained in this project report entitled "**Reliability dimension in review platform:Yelp**" submitted by **Tarun Genwa** (**Roll No.: 150123043**) to the Department of Mathematics, Indian Institute of Technology Guwahati towards partial requirement of **Bachelor of Technology** in Mathematics and Computing has been carried out by him/her under my supervision.

It is also certified that, along with literature survey, **computational implementations have been carried out and empirical analysis has been done** by the student under the project.

Turnitin Similarity: ___%

Guwahati - 781 039
April 2019

(Dr. Ayon Ganguly)
Project Supervisor

ABSTRACT

Reliability and credentiality of a dataset is an important aspect of machine learning and artificial intelligence. The insights made when a dataset is unreliable can lead to pile up of wrong inferences which further lead to incorrect results. Variable selection is a key part of reliability analysis [5]. The aim of our project is to first provide an introduction to Yelp dataset via describing about its big data features and then dive into the various spaces such as user credibility, topic modelling and sentimental analysis.

We have researched and led our analysis in all the aforementioned directions and have derived several insights taken in conjunction and how the platform is affecting the mindset of people and the growth of businesses worldwide. Mostly we have studied the impact of User Reliability, key Business Topics and Rating/Review biases on the Yelp platform to dissect the Reliability of the platform's functioning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Goal

Reliability analysis upon big data systems [5] was our initial goal. Having researched the impact of Yelp platform on revenues and growth of businesses, we have decided to conduct a reliability analysis of the rating system of Yelp. Yelp is an online platform for crowd sourcing business reviews in both local and brand based markets. Yelp businesses covers a wide range of categories to accommodate millions of businesses and have more than 100 million visitors per month on their platform. Figure 1.1 shows the increasing popularity of the Yelp. With such a growing popularity of the platform in metropolitans across the world, its obvious for people to make a quick reference of any business before availing any service or buying something.

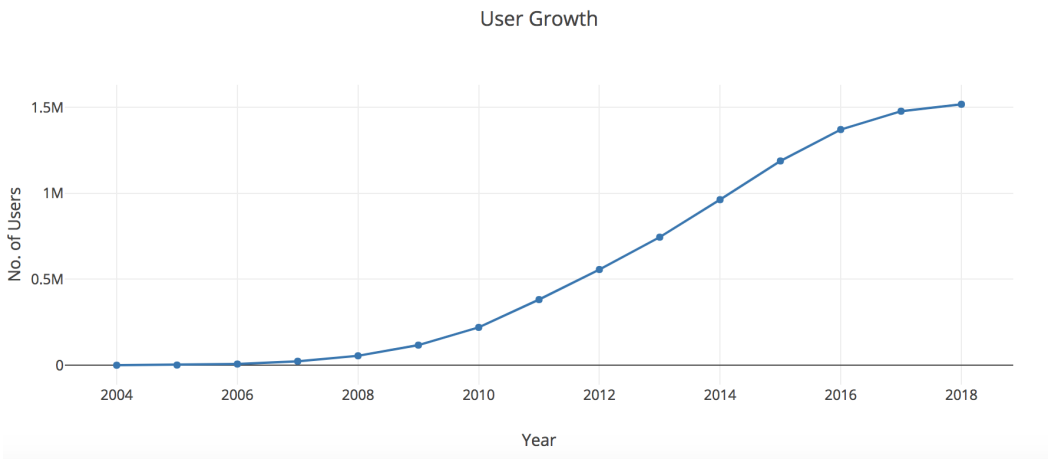

Figure 1.1: User Growth

The motivation behind choosing Yelp dataset as our basis for big data system was mainly due to the effect that it carries on mass opinions and hence comes out to be role player in business growth. Another aspect beneficial to this approach was the availability of a large chunk of company's database in the open source community. The user review data and the business information which Yelp increasingly incorporates at each point of time is enormous and follows the three v's (velocity, volume & variety) of the big data world.

After selecting a platform like Yelp, which gave us around 5.2 million records of user reviews and 174,000 records of businesses sitting in our spark cluster, we decided to explore the data in various ways and draw some interesting conclusions from the same. Then we planned to give users of the platform an eliteness score which could help in giving an extra importance to reviews by these users. Hence, in the first phase of the project we are analysing the reliability of a user based upon a scoring system.

The following sections are the core of our analysis in the project. After taking certain insights from the data model such as star rating distribution

or overall user growth on the platform, we jump onto giving a credibility score to the user using random forest classifier and further classifying the users as elite or non-elite. Further along, we have discussed and used several topic modelling techniques such as NMF, LSA, LSI to obtain the growth factors in a business. After that we have done some sentimental analysis of the review text data and assigned a pseudo rating system which would help in further determining the biases between review text and the rating assigned.

Chapter 2

Exploratory Data Analysis and User Reliability Score

2.1 Exploratory Data Analysis

Let us consider the different collections in our spark cluster from yelp. These all attributes in these collections helps us to analyse the dataset in very different ways and the size of the data is more than a few gigabytes which hinders us from using conventional dataset queries and we have to turn on using big data spark clusters to perform meaningful queries on the dataset. Also the dataset has complex nature due to nested attributes and it is increasing at a very high velocity in terms of reviews as around more than 160 million reviews have been already posted by users on the platform.

```
{
    "review_id": "x7mDIiDB3jEiPGPHOmDzyw"
    "user_id" : "msQe1u7Z_XuqjGoqhB0J5g",
    "business_id" : "iCQpiavjjPzJ5_3gPD5Ebg",
    "stars" : "2",
    "date" : "2011-02-25",
    "text" : "The pizza was okay. Not the best I've had.I
              prefer Biaggio's on Flamingo / Fort Apache.
              The chef there can make a MUCH better NY
              style pizza.",
    "useful" :"0",
    "funny" : "0",
    "cool" : "0"

}
```

Table 2.1: Schema of review.json

```
{
   "business_id" : "Apn5Q_b6Nz61Tq4XzPdf9A",
   "name" : "Minhas Micro Brewery",
   "address" : "1314 44 Avenue NE",
   "city" : "Calgary",
   "state" : "AB",
   "postal_code" : "T2E 6L6",
   "latitude" : "51.0918130155"
   "longitude" : "-114.031674872",
   "stars" : "4",
   "review_count" : "24",
   "is_open" : "1",
   "attributes" : {} 13 items,
   "categories" : "Tours, Breweries, Pizza,
                   Restaurants, Food, Hotels & Travel",
   "hours":{} 6 items

}
```

Table 2.2: Schema of business.json

```
{
   "user_id" : "lzlZwIpuSWXEnNS91wxjHw",
   "name" : "Susan",
   "review_count" : "1",
   "yelping_since": "2015-09-28",
   "friends" : "None",
   "useful" : "0",
   "funny" : "0",
   "cool" : "0",
   "fans" : "0",
   "elite" : "None"
   "average_stars" : "2"
   "compliment_hot" : "0",
   "compliment_more" : "0",
   "compliment_profile" : "0",
   "compliment_cute" : "0",
   "compliment_list" : "0",
   "compliment_note" : "0",
   "compliment_plain" : "0",
   "compliment_cool" : "0",
   "compliment_funny" : "0",
   "compliment_writer" : "0",
   "compliment_photos" : "0"

}
```

Table 2.3: Schema of userreview.json

In order to explore various useful aspects in our reliability analysis, we perform some rudimentary analysis on our Yelp dataset. The pie chart (see Figure 2.1) of of star ratings in the review dataset shows that a majority of reviews are towards higher ratings. However, it also shows that there are lesser median ratings than the lowest ratings.

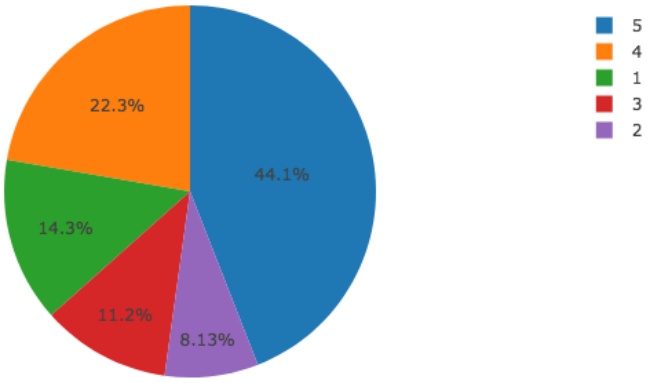

Figure 2.1: Average Star distribution

Figure 2.2 depicts the distribution of the different categories of businesses mentioned in our review dataset. It clearly shows the dominance of businesses like restaurants and food joints with bars & nightlife to have more traction among the community of Yelp. It is somewhat due to the reason that restaurants were the initial target businesses during the launch of the platform.

2.2 User Reliability Score

In order to classify reviews as credible or not, we can create a score for the reviewers which would help in further analysis. In this context, we have assigned eliteness score to each and every user.

Random Forest is basically an ensemble machine learning model which

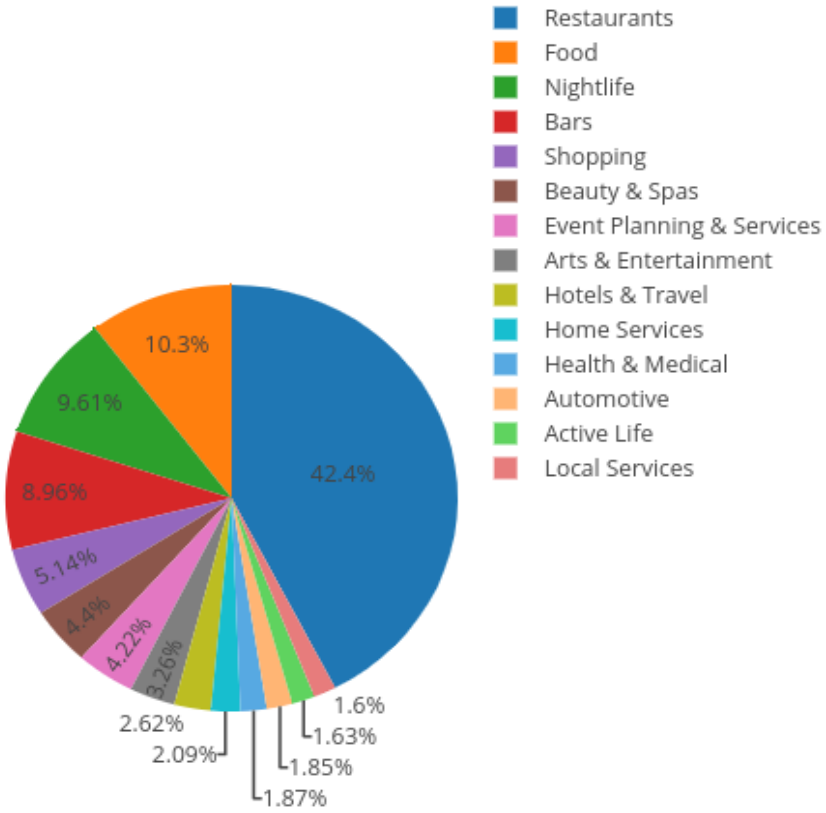

Figure 2.2: Business Distribution

Figure 2.3: Decision Tree Splitting

combines weaker predictive models to make a stronger one. It is a collection of decision trees which in overall reduces the overfitting caused by a single decision tree and gives us a more generalized model. The reason we use random forest algorithm is as follows: the process of bagging where decision

Features	Values
compliment_writer	0.242482
review_count	0.195236
compliment_cool	0.186909
fans	0.184319
compliment_funny	0.071866
compliment_hot	0.049423
compliment_note	0.037891
cool	0.019294
no_friends	0.004900

Features	Values
average_stars	0.004348
compliment_photos	0.001833
compliment_plain	0.000947
compliment_cute	0.000281
funny	0.000159
compliment_list	0.000057
useful	0.000049
compliment_profile	0.000005

Table 2.4: Feature Importance

trees are trained on bootstrapped training sets is parallelizable in Spark. We have used random forest classifier to extract feature importance from our user dataset and also classify them as elite or non-elite. It works by constructing a multitude of decision trees at training time and predicting the average of prediction of actual trees. A typical example is given in Figure 2.3.

We have described the attributes of user record previously, which had an elite label with a boolean value. In order to get the feature importance in the user attributes we take 70% of our user data and run it on random forest classifier. The correcponding weights are given in Table 2.4.

The observed attribute weights can now be used to predict the eliteness score for the whole data set which would be the dot product of feature importance and corresponding feature value in the record. To calculate the accuracy of our classifier model, we compared the boolean elite label and the elite label calculated by the classifier for a given user. The accuracy result came out to be 98%.

A plot of score and corresponding number of users is shown in Figure 2.4. The score distribution calculated after prediction and elite label present in the dataset showcases that only a small percentage of users are elite. From

the actual dataset values it comes out to be 4.62%.

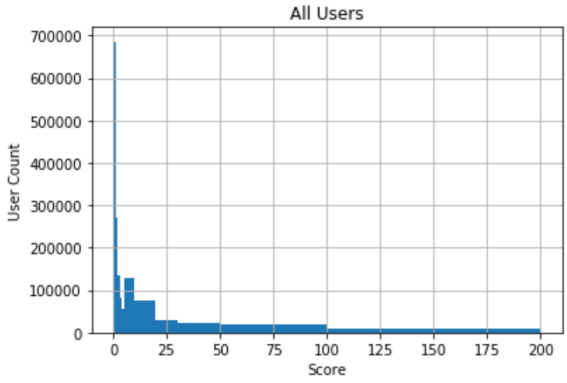

Figure 2.4: All Users Score Distribution

The histogram of scores of elite users is given in Figure 2.5. We can observe that some users with low score are also elite, which means that they lie between the high scoring elite users and high scoring non-elite users.

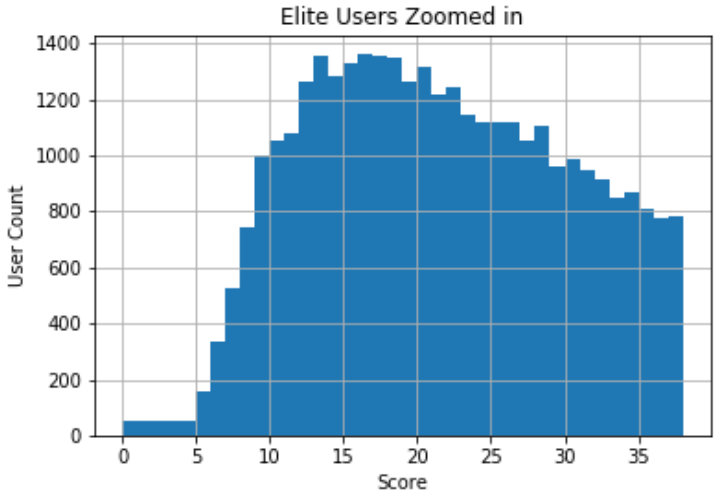

Figure 2.5: Elite Scoring

Similarly we have the distribution of non-elite users and the respective user counts, see Figure 2.6. We can observe the high spikes near score 0

which suggests that a vast majority of users on the platform are non-elite. There are a few number of non-elite users with a considerable high eliteness score which suggests that they have a higher priority than non-elite users but lower priority than all the elite users.

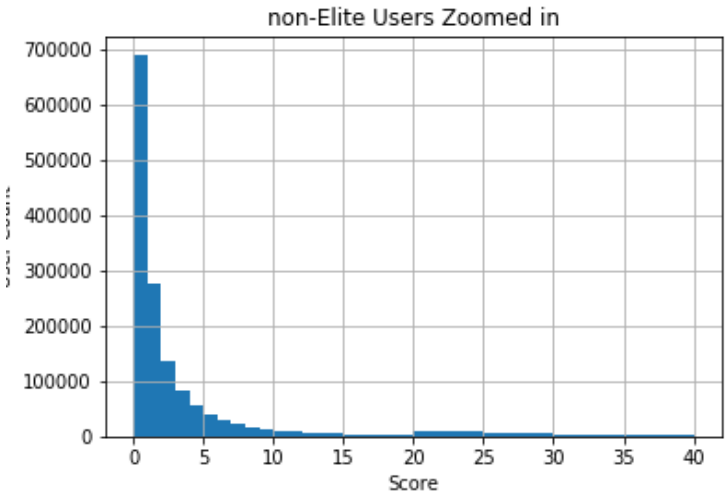

Figure 2.6: Non-Elite Scoring

To measure the performance and validate our classifier result we plot an Receiving Operator Characteristic curve of true positive and false positive rates in Figure 2.7. We can clearly observe that the prediction are highly accurate since the true positive is always high. Area under the curve should be maximum for better results.

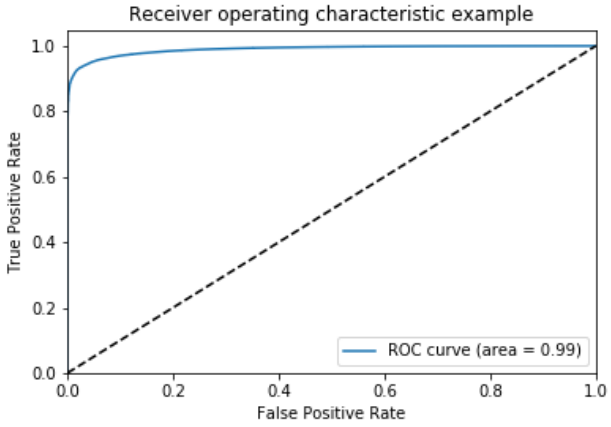

Figure 2.7: Receiving Operator Charactersitic curve

Chapter 3

Topic Modelling

Topic modelling corresponds to discovering abstract 'topics' from a collection of documents. It can be used for information retrieval from unstructured texts in real world datasets using techniques like Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), Non-Negative Matrix Factorization (NMF). Before moving on to the techniques, we need to know what type of pre-processed data is necessary for them. We use Term Frequency-Inverse Document Frequency (TF-IDF) to feed relevant information.

3.1 Term Frequency-Inverse Document Frequency

TF-IDF helps us in finding important words which give an idea about the document. It is product of Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency measures the frequency of a word in a document divided by document length for normalisation.There are some words in a document which have no relevance such as prepositions in a document. So

we need to scale up the rare words and scale down the most frequent ones.

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}} \tag{3.1}$$

$$idf(w) = \log\left(\frac{N}{df_t}\right) \tag{3.2}$$

$$w_{i,j} = tf_{i,j} x \log\left(\frac{N}{df_i}\right) \tag{3.3}$$

where $tf_{i,j}$ = number of occurences of word i in document j, df_i = number of documents containing word i, N = total number of documents. For the dataset, top eight words with respect to relevance of other words in the document in positive and negative reviews are given in Table 3.1. TF-IDF matrix consisting of relevant words are now feeded into LDA, LSI, NMF.

Positive	great, place, service, delicious, best, time, love, order
Negative	order, service, place, time, minutes, restaurant, table, bar

Table 3.1: Top 8 Words for Positive and Negative Reviews

3.2 Latent Semantic Indexing

LSI is a mathematical approach which uses Singular Value Decomposition to scan unstructured and complex data within a document and thus helps in finding hidden (latent) relationships between words to improve upon information indexing. As the underlying mathematical model in LSI is SVD, we can get the corresponding eigenvalues of the important features. Using relative elbow plot, we can determine the required number of topics to be chosen to be used further. LSI also gives us the strength of the reviews. Hence using the relative elbow plot technique we can decide the required number of

topics helpful for further inferences. The elbow plot for the dataset is given in Figure 3.1.

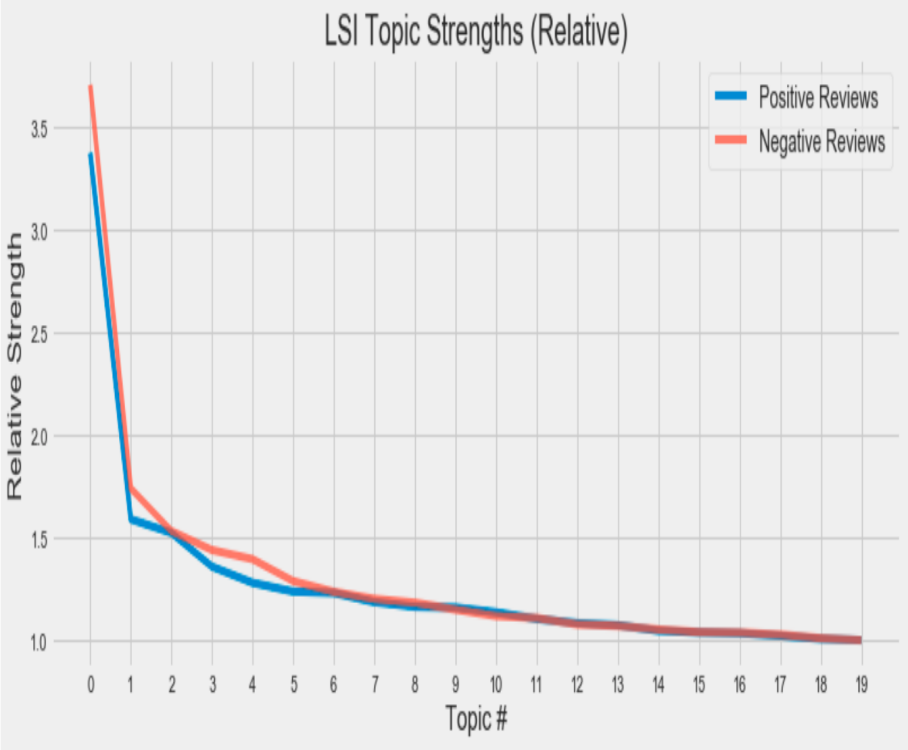

Figure 3.1: Number of Topics Selection

Topic 0	0.296*great + 0.233*place + 0.183*service + 0.135*time + 0.133*delicious + 0.130*order + 0.127*best + 0.126*love + 0.123*pizza + 0.119*nice
Topic 1	-0.859*pizza + 0.259*great + -0.161*crust + 0.143*service + -0.098*sauce + 0.091*atmosphere + -0.080*cheese + -0.070*order + 0.068*place + 0.068*friendly
Topic 2	0.580*great + 0.348*pizza + 0.202*service + -0.197*"order + -0.189*chicken + 0.133*atmosphere + 0.124*place + 0.106*beer + 0.105*staf" + 0.098*friendly
Topic 3	-0.581*"burger + -0.289*beer + -0.250*fries + 0.249*sushi + -0.188*"bar + 0.150*thai + -0.147*selection + 0.130*service + 0.118*restaurant + -0.115*cheese
Topic 4	-0.420*place + -0.417*love + 0.246*great + 0.222*service + -0.219*"best" + -0.215*burger + -0.187*sushi + 0.181*excellent + -0.166*cleveland + -0.114*favorite

Table 3.2: Top 5 Positive Topics using LSI

Topic 0	0.229*order + 0.192*service + 0.187*place + 0.164*time + 0.154*minutes + 0.126*table + 0.120*restaurant + 0.116*server + 0.111*"bar" + 0.101*chicken
Topic 1	-0.323*minutes + 0.232*chicken + 0.225*pizza + -0.187*table + -0.159*waited + 0.156*sauce + -0.140*wait + -0.127*server + 0.123*cheese + -0.122*waiting
Topic 2	-0.898*pizza + 0.185*burger + 0.107*chicken + -0.092*crust + 0.085*fries + 0.064*rice + 0.063*steak + -0.055*pepperoni + 0.052*sushi + 0.050*meal
Topic 3	0.408*burger + 0.356*order + -0.294*place + -0.292*service + 0.201*minutes + 0.165*fries + -0.149*sushi + 0.147*pizza + 0.135*chicken + 0.112*salad
Topic 4	-0.686*burger + 0.283*chicken + -0.201*fries + -0.199*service + -0.158*beer + -0.132*place + 0.131*rice + -0.128*bar + 0.121*"order + -0.097*slow

Table 3.3: Top 5 Negative Topics using LSI

3.3 Non-Negative Matrix Factorization

Linear algebraic model which reduces high dimensional vectors into a low-dimensionality representaion being similar to PCA,NMF uses non negative vectors and forces the coefficients to also be non-negative.

Topic 0	1.127*order, 0.980*chicken, 0.674*time, 0.602*delicious, 0.593*menu, 0.590*sauce
Topic 1	3.124*pizza, 0.569*crust, 0.241*sauce, 0.228*best, 0.214*pepperoni, 0.204*cheese
Topic 2	3.122*great, 0.684*service, 0.670*beer, 0.582*atmosphere, 0.535*bar, 0.523*selection
Topic 3	2.390*burger, 0.996*fries, 0.469*beer, 0.243*cheese, 0.210*selection, 0.193*bar
Topic 4	1.938*place, 1.403*love, 1.000*best, 0.808*cleveland, 0.742*sushi, 0.535*favorite

Table 3.4: Top 5 Positive Topics using NMF

Topic 0	1.962*place, 0.802*bar, 0.578*beer, 0.566*restaurant, 0.559*sushi, 0.556*great
Topic 1	1.270*chicken, 0.792*order, 0.656*salad, 0.647*sauce, 0.498*cheese, 0.485*tasted
Topic 2	3.027*pizza, 0.324*crust, 0.242*cheese, 0.196*order, 0.189*pepperoni, 0.189*sauce
Topic 3	1.586*minutes, 1.129*order, 1.079*table, 0.796*wait, 0.791*server, 0.752*waited
Topic 4	2.716*burger, 0.909*fries, 0.362*medium, 0.353*order, 0.229*beer, 0.225*bun

Table 3.5: Top 5 Negative Topics using NMF

3.4 Latent Dirichlet Allocation

Here each document is used to describe distribution of topics and each topic described by distribution of words. LDA assumes a fixed number of topics in which each topic represents a set of words. A mapping is established from all documents to each topic in such a way that most of the words in each document are captured by the corresponding topics.

Topic 0	0.015*great + 0.011*order + 0.008*place + 0.007*delicious + 0.007*chicken + 0.007*service + 0.006*restaurant
Topic 1	0.023*pizza + 0.011*place + 0.010*sushi + 0.009*"order" + ' 0.009*best + 0.008*great + 0.006*time
Topic 2	0.026*great + 0.013*place + 0.011*service + 0.011*burger + 0.009*"bar + 0.007*nice + 0.007staff
Topic 3	0.020*great + 0.018*place + 0.015*service + 0.012*time + 0.010*order + 0.009*chicken + 0.009*"best
Topic 4	0.019*place + 0.007*great + 0.006*menu + 0.006*bar + 0.005*cleveland + 0.005*restaurant + 0.005*pretty

Table 3.6: Top 5 Positive Topics using LDA

Topic 0	0.017*order + 0.014*minutes + 0.013*service + 0.012*time + 0.011*server + 0.011*table + 0.009*place
Topic 1	0.019*order + 0.014*place + 0.013*service + 0.008*time + 0.008*waitress + 0.008*asked + 0.007*minutes
Topic 2	0.013*place + 0.010*order + 0.008*pizza + 0.007*"service" + ' 0.006*menu + 0.006*sauce + 0.006*better
Topic 3	0.015*order + 0.010*chicken + 0.009*time + 0.007*place + 0.007*cheese + 0.006*service + 0.006*people
Topic 4	0.012*place + 0.011*service + 0.008*bar + 0.008*order + 0.007*beer + 0.007*time + 0.006*better

Table 3.7: Top 5 Negative Topics using LDA

3.5 Comparison between LSI, LDA, and NMF

Table 3.2 - Table 3.7 provides the basis for the following comparison. LSI gave us topics but had given weights to the words in the range of [-1,1] from which we cant infer the relative importance of words in a topic. So we tried NMF and LDA which gives positive weights to the words in a topic. Topics given by LDA had a lot of overlapping of words but NMF on the other hand had almost no overlapping thus providing more discrete topics and more relevant information to guide the business owners.

Chapter 4

Sentiment Analysis

Sentiment Analysis refers to contextual meaning of text. In our project we have been able to utilise this technique to extract subjective information from our source material which is predominantly review text. The goal in this regard is to understand the social sentiment towards a brand or business from the platform users and draw some key insights which value in regard to improvising and sustaining the business in the long run.

Our aim with sentimental analysis is to pick out the emotion of the review text and analyse with its corresponding rating as if to analyse the bias between both of them. With further extrapolation, after applying naive bayes and similar strategies, it gives a pathway to build a pseudo rating system which would be able to generate rating based upon random textual input. The pseudo rating system if trained properly can than be used to cross check any review in the dataset and point out the degree of biases in the same.

To begin with sentimental analysis in our textual review data we need to use some Natural Language Processing(NLP) preprocessing techniques to clean our data before analysis. We start by removing stop words in the dataset. Stopwords are the ones that are irrelevant in the dataset and we

would like to exclude them from our analysis. These stopwords needs to be mentioned manually before running any NLP algorithm. Some libraries support automatic removal of stop words with some predefined dataset of these words.

Naive Bayes for prediction of positive or negative review gave us a f1-score of 0.97. Naive Bayes for prediction of star rating when processed only upon the review text gave us a f1-score of 0.63 which is evident from the fact that the stars are just not dependent on the review text rather on many parameters like cool, funny, useful, geographical locations and much more. Even vectors prepared in the NMF technique which can be used for review similarity can be used for cross validation of a particular review.

Chapter 5

Conclusion

In our two semester long project, we have tried to analyze Yelp dataset in different ways and have managed to carry out some interesting insights about the platform and crowd sourced review platform culture in general. We started with exploratory data analysis and mainly with user reliability analysis where we demonstrated that a large population of our platform users are non-elite and those with the elite tag have the implicit credibility in their reviews or comments. This classification of users helps the platform in fake review filtering by giving weights to the reviewing party. Next, we introduced topic modelling, the part which led us to analyze the review text. After applying techniques such as LSI, LDA, NMF, we were able to comparatively identify topics and assigned them into two categories, *viz.*, positive and negative topics. These topics if conveyed to the business owners on the platform may help them in realizing the pros and cons of their establishments functioning and in turn helping them to improve their businesses. At the end, we conducted sentimental analysis to determine the bias between the users ratings and the emotion carried in their review text. It helps in segregation of reviews where a users assigned rating does not carry the meaning associated

in the text due to miss judgment of rating index.

After our analysis, we would like to quote some key facts or findings from previous researchers on the Yelp platform. They say that about 10 million reviews are annually filtered out as fake by Yelps filtering algorithm which amounts to a quarter of the total submitted reviews. This is a massive integrity check on the first level for the platform. It is also found out that having an extra star on Yelp causes the business to rise by 5 to 10 percent emphasizing a direct connection between the Yelp ratings and business bottom line.

Now overall, after conducting our analysis in three different directions, we would like to conclude our findings and try to draw some meaning out of all of them when seen in conjunction. A increased percentage of reviews by reliable users will benefit the businesses which are actually doing good work and people would be less likely to be disguised by the false reviews. The businesses which would focus on the topics found out via Topic Modelling can achieve major gains by improving in those scenarios, thus the platform would prove helpful for those type of businesses. At last, the method of sentimental analysis to give a pseudo rating and cross checking the positivity or negativity of a review with its' indexed rating will definitely keep the sanity check of the reviews published on the platform.

Since proving that whether the crowd-sourcing review platform Yelp is reliable or not is not a binary, it is mostly a distribution of pros and cons given to it in various sections. The ability of malicious and ill-intended reviews to creep in the system is cruel but the measures to keep it sanitized has been rigorous. We, from our analysis, can infer that if in-together users are more reliable on the platform, with the businesses getting a feedback of their performance on a daily basis and also the actions of users are kept a

check on, a pretty efficient and not so vulnerable system of crowd-sourced review and ratings can be maintained for a prolonged period.

In conclusion, we would like to say that there are a hundred different ways this huge dataset can be utilized and analyzed to ascertain more vulnerabilities and focal points of the platform which will help in keeping it a reliable source of making decisions.

Bibliography

[1] James Huang, Stephanie Rogers, and Eunkwang Joo. Improving restaurants by extracting subtopics from Yelp reviews. *iConference 2014 (Social Media Expo)*, 2014.

[2] Chen Wen Li and Jin Zhang. Prediction of Yelp review star rating using sentiment analysis. 2014.

[3] Kyong Jin Shim Koo Ping Shung Maruthi Prithivirajan, Vivian Lai. Analysis of star ratings in consumer reviews. *IEEE International Conference on Big Data*, 2015.

[4] Y. Paule Sun. Spatial analysis of users-generated ratings of Yelp venues. *J.D.G. Open geospatial data, softw. stand.*, 2017.

[5] William Q. Meeker Yili Hong, Man Zhang. Big data and reliability applications: The complexity dimension. *Journal of Quality Technology*, 2018.