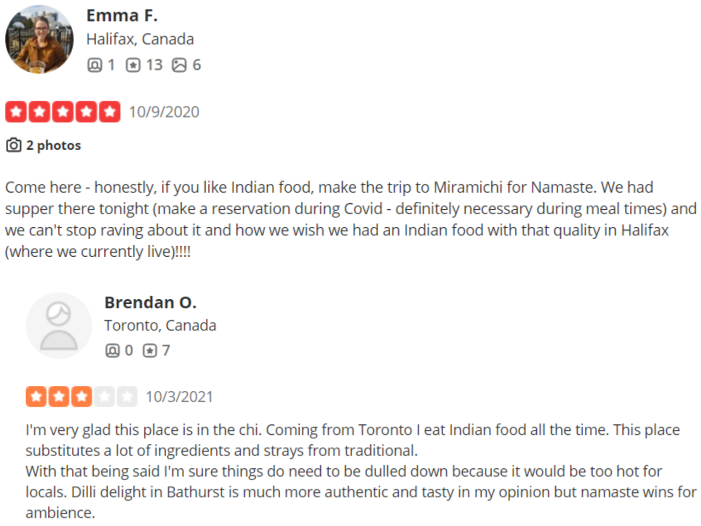
Predicting user ratings in a Yelp network

Parth Pandya, Pratyush Thakur, Eknath Das

Abstract- Yelp network is a platform where users can find many small businesses, restaurants and service companies.It is network of businesses and users where some users share their experiences about the places while others can take decisions based on their experiences. Users share their experiences by writing textual reviews and giving ratings in terms of starts about the business or service. It has been observed in some examples that two users might have written equally positive or negative reviews but there is a huge difference between the ratings that they give. So we can say that to figure out correct details about the place one needs to read all the textual reviews and not just be dependent on the rating stars which is a very tedious task. This paper gives a method by which we can predict the stars based on all the textual reviews.

I. Introduction

Textual reviews are very subjective in nature. It is essential to summarize this review to make the correct decision about the place. In figure 1.a there are two different persons reviewing the same restaurant. They both had written positive reviews for the place but they have given star ratings very differently. This can happen for various reasons like people's personality or personal bias.This is why review summarization is very important to give correct information to other users otherwise user have to read all the tons of reviews to make correct judgment about the place.



Emma F.
Halifax, Canada
1 13 6

★★★★★ 10/9/2020

2 photos

Come here - honestly, if you like Indian food, make the trip to Miramichi for Namaste. We had supper there tonight (make a reservation during Covid - definitely necessary during meal times) and we can't stop raving about it and how we wish we had an Indian food with that quality in Halifax (where we currently live)!!!!

Brendan O.
Toronto, Canada
0 7

★★★☆☆ 10/3/2021

I'm very glad this place is in the chi. Coming from Toronto I eat Indian food all the time. This place substitutes a lot of ingredients and strays from traditional.
With that being said I'm sure things do need to be dulled down because it would be too hot for locals. Dilli delight in Bathurst is much more authentic and tasty in my opinion but namaste wins for ambience.

Problem statement here is to predict the rating in terms of stars based on the tons of textual reviews. Predicting a rating that a user will give to a business or service is very useful in building a recommendation system. Yelp has a five star rating system so if we can predict earlier what rating will this user give to the particular restaurant then the system can provide very helpful recommendations.

We can go ahead with three different approaches for the given problem statements. We can analyze the similarity between the products that are being reviewed and then based on the rating in terms of stars given to the similar product by the same user after writing the review, we can predict the rating for the given product. Another approach is that we find the similarity between the user's friends and user and then based on the review and rating given by the friends network of the user we can predict what rating the user will give.Third approach suggest the analysis of all the past textual review for that product and after performing sentiment analysis we can predict the rating for the product.

II. Related work

There are many research papers published in the area of textual analysis, text summarization or recommendation systems. Nowadays the amount of data available for research in this field is very large.Yelp network itself has provided a very big dataset of their network for Yelp rating prediction challenge.

In this field among all the papers two major approaches have been identified.In both the approaches one of the major tasks to solve this problem is to extract the important features from textual data..

Yand, Yuan and Zhang have provided the collaborative approach for predicting star rating in the yelp network. In which they have considered the influence of the friends of users also as one of the significant features.They have used standard latent factor models for predicting the rating.

$$R_{u,i} = \mu + a_u + b_i + q_i^T p_u$$

Where, Ru,i is the prediction of the rating for item i by user u. μ is a global offset parameter. au and b i are user and item biases respectively. pu and qi are user and item factors. Fig 1.b represents the entire friendship network of the users in yelp platform.

Almashraee and Paschke have provided a feature extraction technique based on Semantic sentiment analysis.They have used naive bayes classifier for opinion mining. Opinion mining is the sub field of text mining. This method gives the most significant features for sentiment analysis. It gives most relevant data through which the model can easily decide the opinion from the text.

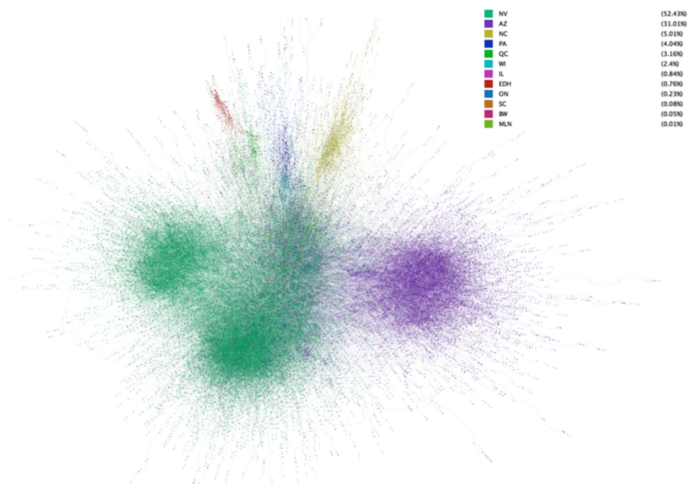


Fig 1.b Friendship network color coded by location

III. Algorithm

The proposed approach here is divided into three parts. First part of the approach is Data loading and cleaning. Yelp dataset is very huge so loading preprocessing of the data and converting it into review user data is part one of the approach.Second part of the approach is feature extraction from the data.Textual data contains a lot of unnecessary data so feature extraction from that data is challenging task. After selecting the feature, the third part of the approach is about selecting the models and comparing it with other models.

A. Data

We have used the Yelp challenge dataset. This dataset contains 269231 users, for this experiment we have used only restaurant business data into consideration so this dataset contains 21892 restaurants, 990627 reviews,check ins and tips.

1. Data Preprocessing

After loading the dataset we need to organize the data in a way that it can be directly applicable to the model and have a minimum amount of noise. There is two type of data here one contains the data of every user and it's reviews while the other contains the

We have analyzed the length of the reviews of the users and we find out that the most useful and important reviews length is in the range of 50-100. So instead of working with a very large dataset we have decided to work with only reviews with this length which includes almost every useful review. Fig 3.a is a visual representation of our analysis. In figure we can see that almost all the reviews are in the range of 50-100 only. Fig 3.b is a visual representation of distribution of rating stars.

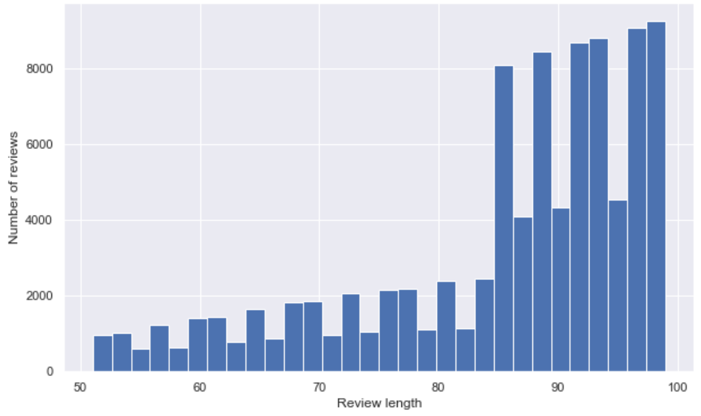


Fig 3.a

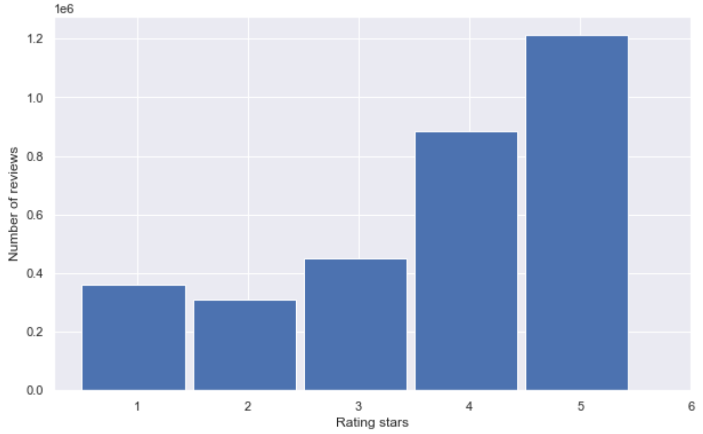


Fig 3.b

B. Preparing and preprocessing the review

In preparing and preprocessing the review nltk library was used to get a list of stopwords. 3 steps were taken

1. Processing reviews to extract characters in between a-z and A-Z only
2. Removing the stopwords
3. Reviews between length 50 to 100 are taken
   We were able to segregate 95017 reviews

C. Model

1. **Naive Bayes model((TF-IDF features))**

a. Dataset preparation :
   We would be training a text dataset which would be having list of list prepared from preprocessing as in the previous step.Here each list element would be a review prepared after preprocessing.

b. Predicting stars from the text :

The TfidfVectorizer() method from sklearn library is used to generate features (TF-IDF) from the text document provided. For this model we used 'Nearest Neighbors', 'Multinomial Naive Bayes', 'Logistic Regression'. Accuracy obtained is shown in the table in the result section.

2. LDA Model

a. Data preparation :
The dataset which is prepared in the preprocessing step is further processed.

b. Topic distribution matrix :
Generated a matrix of topic probabilities for each document in the matrix , so here it would be taking an input corpus and keeping track of topic distribution of each review that can be used as the features predict the number of stars.

c. LDA data frames :
Using LDA (Latent Dirichlet allocation) we would be generating topics and storing topics for each star rating separately.
These lists would be concatenated with the
corresponding star and this would be the training data

d. Predicting stars from the text :
'Nearest Neighbors', 'Multinomial Naive Bayes', 'Logistic Regression', 'LDA', 'QDA', 'Random Forest', 'AdaBoost' are the models used The results found are listed in the table in the result section

3. TF-IDF features and sentiment

1. Sentiment column :
To get more accurate results we further added a column to the features called as sentiment where we would be having value 1 when rating > 3.5 and 0 when <=3.5.
2. Dataset preparation:
We would be using the TF-IDF features along with the sentiment.
3. Results obtained are the best as reported in the table

4. LDA and sentiment

Here we concatenate the sentiments along with the LDA features generated in the model 2.

IV Results

We have compared four models with different classifiers such as nearest neighbor, multinomial naive bayes, logistic regression, LDA, QDA ,Random forest and Adaboost. After performing this experiment for comparative study we came to find that logistic regression with sentiment results model gives the best accuracy. Here we have put the results for each model table. Here, we have defined the terminologies used in the table.

Model 1: Naive Bayes model((TF-IDF features))
Model 2: LDA Model
Model 3: TF-IDF features and sentiment
Model 4: LDA and sentiment

V. Conclusion

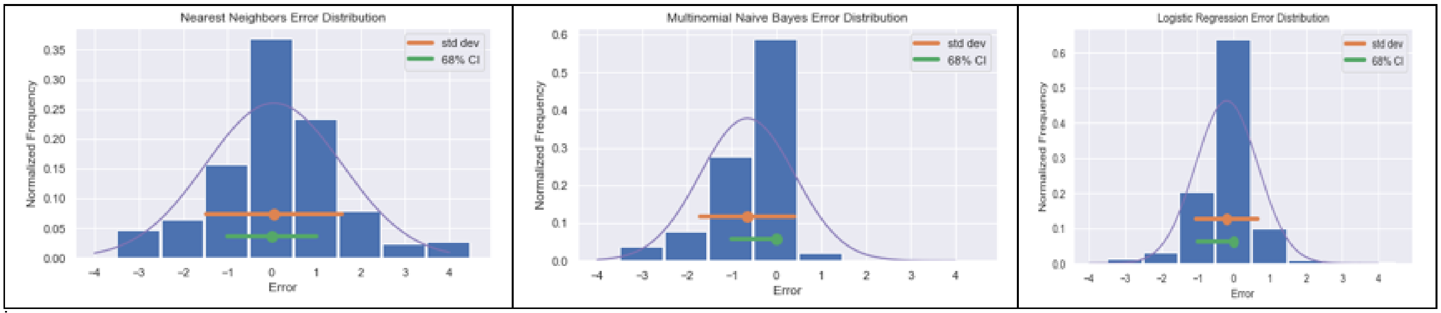
We were able to successfully compare various models in terms of how the features are being provided to the system. Firstly ,we used vectorized version of the features using the TfidfVectorizer then the features generated using the topic distribution via latent dirichlet allocation (LDA) and then contained both using the sentiments .We were able to analyze that the best accuracy was obtained using the TF-IDF features in the logistic regression method.

Accuracy	Nearest neighbor	Multinomial naive bayes	Logistic Regression	LDA	QDA	Random Forest	Adaboost
Model_1	0.35936	0.565085	0.632642	-	-	-	-
Model_2	0.408292	0.515942	0.515679	0.514417	0.453383	0.477586	0.514522
Model_3	0.727297	0.867095	**0.900295**	-	-	-	-
Model_4	0.499632	0.515942	0.574187	0.571556	0.555404	0.539882	0.567873

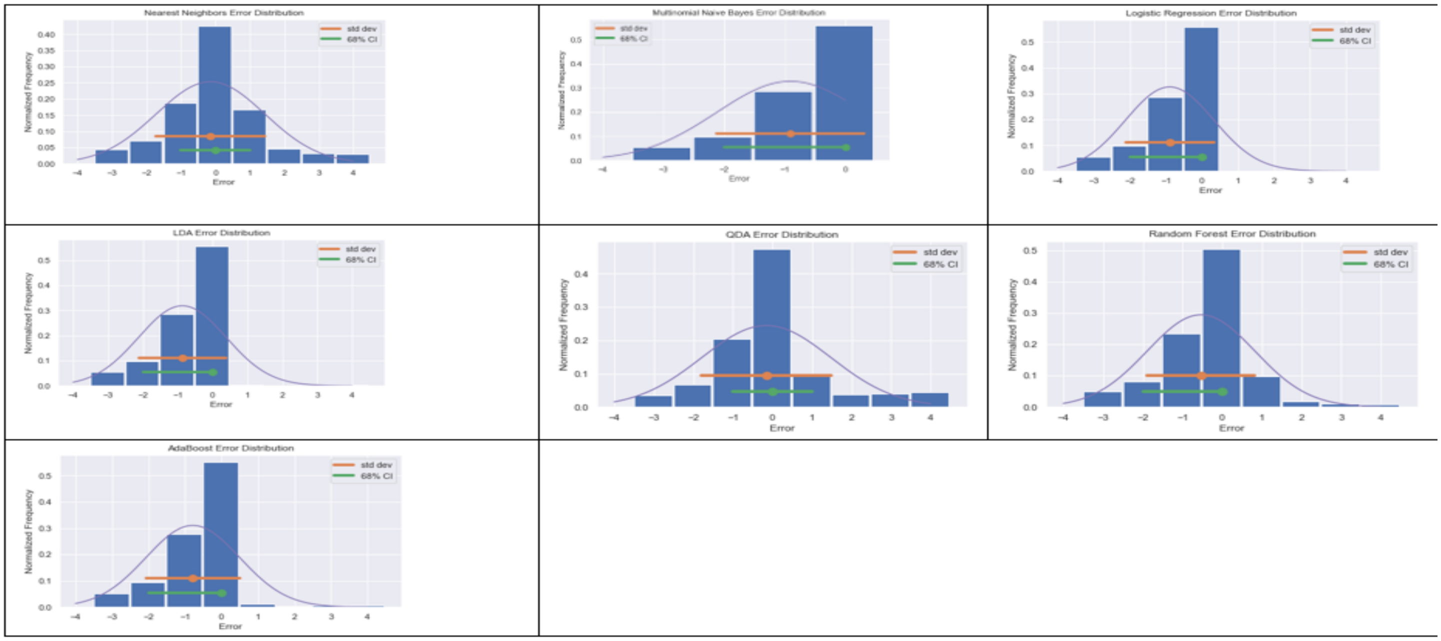
Table 1. Result

This is the visual representation of **error distribution** in each classifier in a particular model.

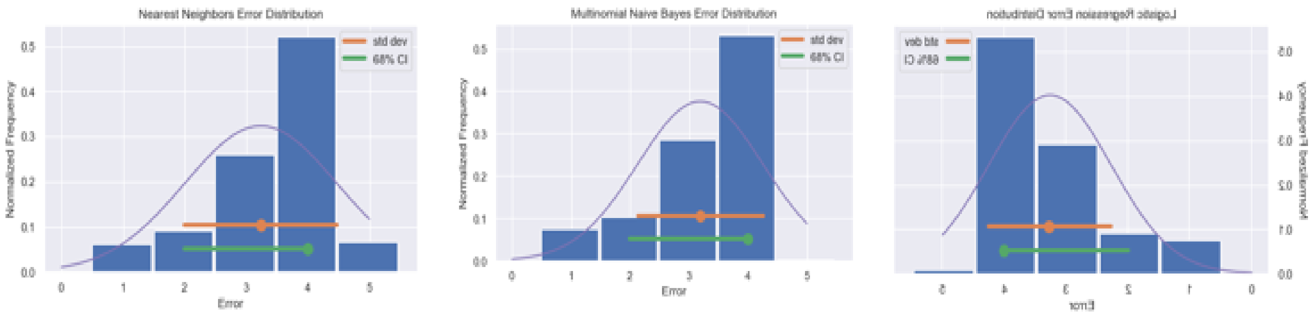
Model 1 :



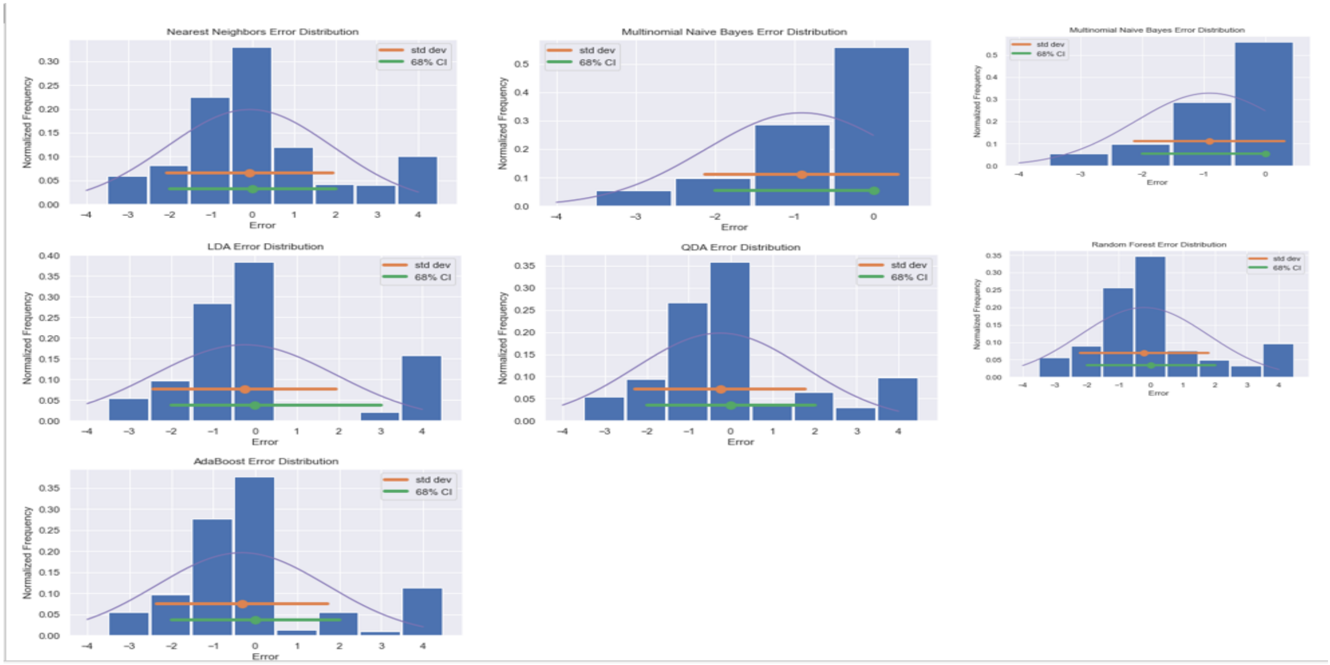
Model 2:



Model 3:



Model 4:



VI. Reference

1. Predicting a Business' Star in Yelp from Its Reviews' Text Alone Mingming Fan,Maryam Khademi
2. Feature Extraction based on Semantic Sentiment Analysis Mohammed Almashraee and Adrian Paschke
3. Predicting Yelp Ratings Using User Friendship Network Information Wenqing Yang (wenqing), Yuan Yuan (yuan125), Nan Zhang (nanz)

EFFORT DISTRIBUTION :

1) Pratyush(2020H1030121P) : Prepared the topic matrix using the LDA features ,further used the scikit learn library on various models like lda(),qda(),multinomial naive bayes,knn, logistic regression. Used them for training the models and further generating the results in terms of accuracies , precisions etc.Features for the sentiments were also prepared and used along with various models.Prepared final report on assignment 2.

2)Parth Pandya(2020H1120287P) : Preprocessed the dataset to segregate reviews by using the business data and extracting from the business_yelp.csv given on the official website.

Further the reviews were prepared by extracting only text data and removing the stopwords from it. Also the features were distributed and concatenated with the corresponding star rating to get the training and the test set.Prepared final report on assignment 2.

3)Eknath dhas (2020H1030140P) : Read research papers and understood them , further prepared the part 1 assignment write up.Explored the areas which can be worked upon.