

Homework 6: Prediction of rating from Yelp review text

Grover, Karan & Arora, Pragya & Ghai, Piyush

{grover.120, arora.170, ghai.8}@osu.edu

November 30, 2016

1 PROBLEM STATEMENT

1.1 ABOUT YELP

Yelp [1] a famous website as well as a mobile app which publishes crowd sourced reviews about food joints and businesses. It also has a division which handles online reservations for restaurants. **Yelp Dataset Challenge**[2] is a publicly open contest sponsored by Yelp, in which the participants are challenged to use Yelp's data in an innovative way.

1.2 OUR MISSION

The Yelp dataset downloaded from Yelp dataset challenge website is huge and consists of over 9M reviews by roughly 687k users for over 86k businesses [2]. For this project, we chose to predict a review's rating based on the review text. The rating will be done on a scale of 1-5, where 1 stands for awful and 5 stands for excellent. We built multiple models and accessed which models would fit our use case the best. This is explained in more depth in the later sections of this report.

1.3 ABOUT THE DATASET

The Yelp Dataset consists of several files in JSON format of the data. The main files as per our use-case were the *yelp_academic_dataset_business.json* & *yelp_academic_dataset_review.json*. The two data files were **2.13 GB** & **73.6MB** in size respectively. The data representation in the for *academic_dataset_business* is as follows :

```
1 {
2     "type": "business",
3     "business_id": (encrypted business id),
4     "name": (business name),
5     "neighborhoods": [(hood names)],
6     "full_address": (localized address),
7     "city": (city),
8     "state": (state),
9     "latitude": latitude,
10    "longitude": longitude,
11    "stars": (star rating, rounded to half-stars),
12    "review_count": review count,
13    "categories": [(localized category names)]
14    "open": True / False (corresponds to closed, not business hours),
15    "hours": {
16        (day_of_week): {
17            "open": (HH:MM),
18            "close": (HH:MM)
19        },
20        ...
21    },
22    "attributes": {
23        (attribute_name): (attribute_value),
24    },
25 }
```

The data representation in the for *academic_dataset_review* is as follows :

```

1 {   "type": "review",
2     "business_id": (encrypted business id),
3     "user_id": (encrypted user id),
4     "stars": (star rating, rounded to half-stars),
5     "text": (review text),
6     "date": (date, formatted like "2012-03-14"),
7     "votes": {(vote type): (count)},
8 }

```

The given datasets were first converted and exported into csv formats using a simple python script. The python script is a part of the **PreProcessing1.py** python file.

2 EXPLORATORY ANALYSIS OF YELP DATASET

The primary features of interest to us in the given dataset are **reviews** by the user, the **stars** associated with a review and the business establishment the reviews belong to. After merging the business dataset and the review dataset, we analysed which business unit the maximum reviews belonged to. Table 2.1 lists out randomly sampled 10 rows from the dataset. In figure 2.1, the x-axis was chosen as numeric in order to display a cleaner graph, but the sharp spike corresponds to *restaurant* as the main category. The total review count is close to 1.6M for this category. Next we thought of finding the geographical split of the reviews. The Yelp dataset consists of data collected from 29 different states. Table 2.2 shows the split of review count over all business establishments sorted by the top 7 states. As we can see, the maximum reviews come from Nevada, followed by Arizona, both of which are over 1M in number. After Arizona, the next highest is North Carolina, which is nearly one-tenth lesser than that of Arizona. This could perhaps be indicative of a very strong presence of Yelp in business establishments in NV and AZ as opposed to other states.

Since there are so many reviews for restaurant, we decided to focus on the subset of reviews for restaurants (the subset is obtained in the pre-processing step) for further analysis. Next we observed the variation of average rating stars against the *length of the reviews (in characters)*. The average rating varies quite a lot for reviews with higher number of characters, indicating that the polarity of reviews fluctuates a lot more from one length to another. There is less fluctuation in the average rating of reviews from length 50 to 500 characters. Hence we chose to subset the reviews based on minimum and maximum length thresholds. Figure 2.2 indicates this observed trend. Another motive for reducing the subset for reviews is based from inferences drawn from Figure 2.3. From this we can see that the total number of length >700 is very less as compared to the overall size of the review corpus. Thus combining the analysis from this and the previous figure, we subsetting the number of reviews.

Next we observed the distribution of reviews with the star rating, i.e. what review was given which star by a user. From Figure 2.4, we can see that there is a skewed distribution of reviews in terms of star ratings they have received. A majority of reviews have a 5 star or a 4 star rating, while the count for 1 star is the lowest. This will later on form a basis for us to create training datasets based on data sampled from each star rating, in order to ensure even representation in the training corpus.

2.1 WORD CLOUD

Given the distribution of reviews, we decided to capture the common sentiment for all the reviews individually. We did this by plotting a wordcloud of review texts grouped by their star rating. To do this, we preprocessed the review texts to remove all the common stop words, tokenized the words and then plotted them. Figures 2.5, 2.6, 2.7, 2.8, 2.9 represent this pictorially.

Table 2.1: Categories and sub categories a review belongs to in the Yelp business dataset

ReviewId	Categories
QPTBica7UkkrT7xEQskBww	[Nightlife, Bars, American (Traditional)]
O93XUShpjG7X_VioaN0ViQ	[Local Services, Mobile Phone Repair]
5zvonAPvVG17Aid1MZYFGw	[Food, Coffee & Tea]
UrA_eN4seEdDLZR-gIrG4g	[Skin Care, Weight Loss Centers]
vZSNSfzb7XDMpBXr_PEX1w	[Music Venues, Arts & Entertainment]
CEO6MmROql9ZFHKtQ1FXow	[Bars, American (Traditional), Food]
lH-jQdLKKGcG-3DzgMgLQA	[Thai, Restaurants]
ctbEMtUouLPzf2F25DQamw	[Tires, Automotive]
iESG5qEoUNYJ4lOOv0337g	[Burgers, American (New), Restaurants]
w0LtO11Ylu-7wnzKs1E9QA	[Sushi Bars, Asian Fusion, Restaurants]

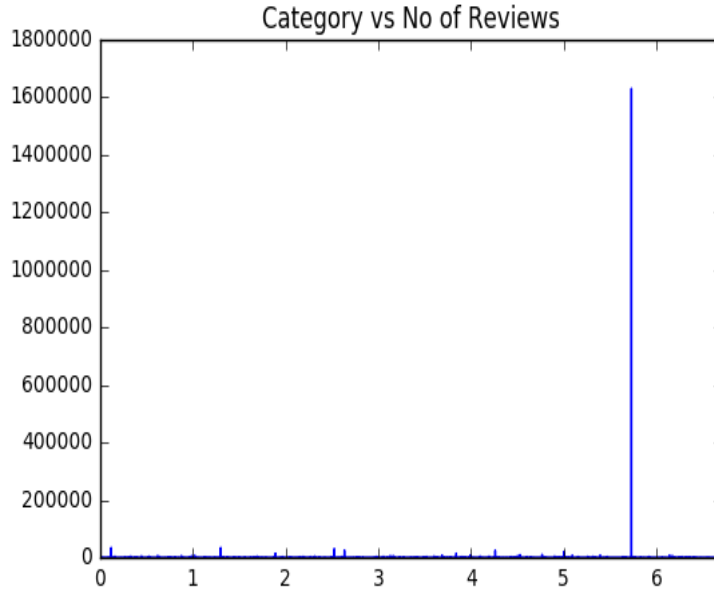


Figure 2.1: Review count v/s category they belong to

Table 2.2: Review counts for different states

State	Total Reviews
NV	1154799
AZ	1029103
NC	165625
PA	111542
QC	88046
WI	69917
EDH	30003

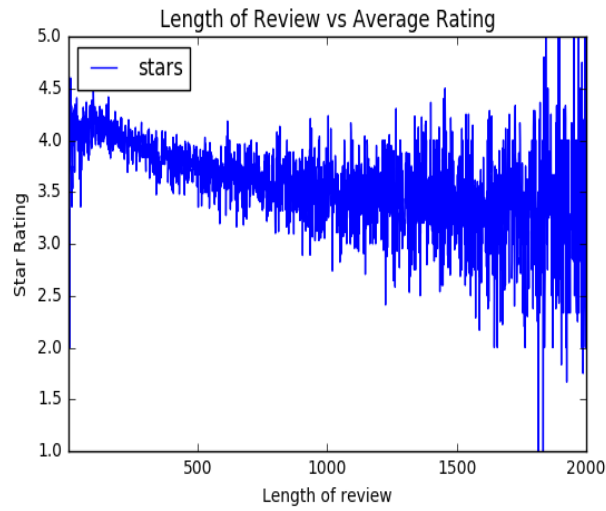


Figure 2.2: Review count v/s category they belong to

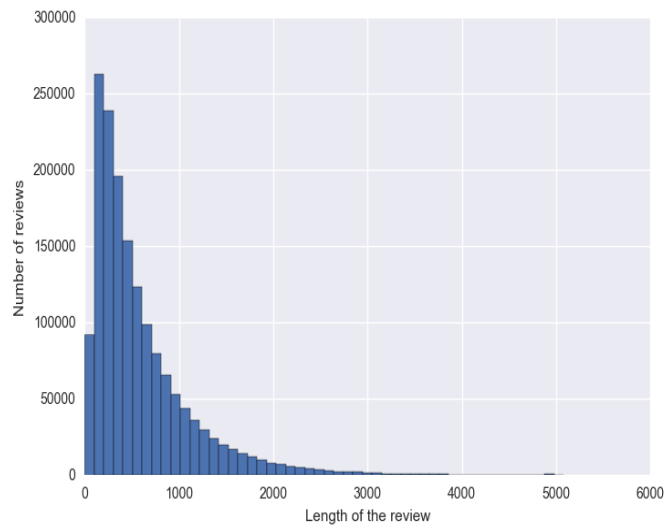


Figure 2.3: Length of a review v/s count of reviews

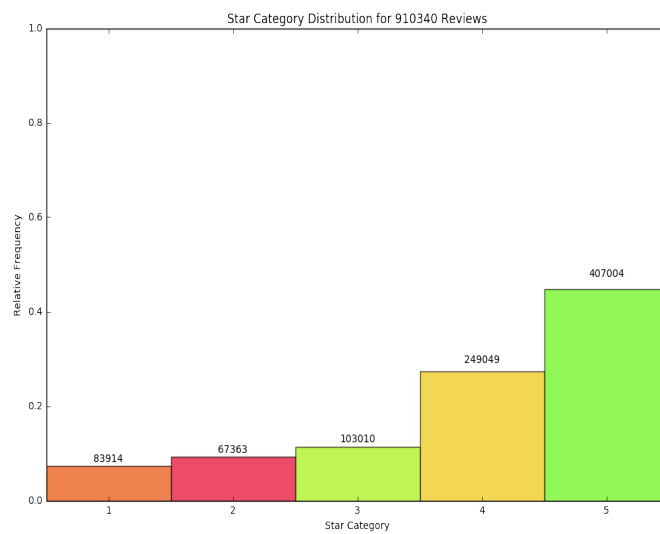
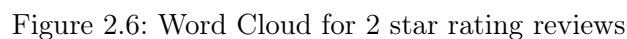
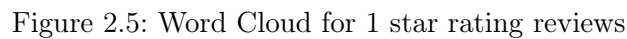


Figure 2.4: Total reviews for a given star rating



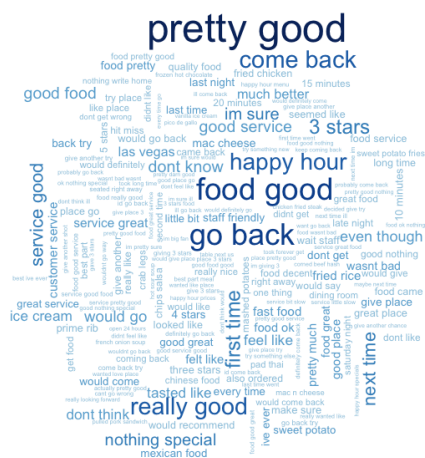




Figure 2.9: Word Cloud for 5 star rating reviews

3 PROGRAM DESCRIPTION

The code has been broken down into the following parts :

1. **Pre-Processing 1**

In pre-processing 1, Yelp provided JSON files were converted into the CSV format. The required CSV files were loaded as Python data-frames and then merged, filtered and stored locally for later usage.

2. **Pre-Processing 2**

In pre-processing 2, the review text is cleaned up, grouped and stored according to the rating of the reviews, which further gets partitioned into training and testing dataset.

3. **Models**

We implemented a total of 8 models and each model has been separated into a different python file. Following are the models which were implemented and compared:

- **Baseline model**
- **TFIDF model**
- **Bag of words model**
- **Bigram model**
- **Trigram model**
- **Bi and Tri-gram model**
- **LDA Model**
- **LDA + Sentiment layer model**

4 MODEL DESCRIPTION

We built several models which generated different forms of features of review text to help predict the sentiment:

1. **Baseline model**

In the baseline model, the average rating across all reviews is assigned as the predicted rating for each review. The average rating of all reviews was found to be 3.9 which was rounded off to 4 and used as baseline prediction. This was used as a simple benchmark for all the other models we applied to this dataset.

2. **TFIDF model**

In TFIDF or term frequency-inverse document frequency, each review text is represented as a vector having dimension equal to the number of words in the dictionary and each dimension representing a weight of importance of the word in the review text. This TFIDF vector representations of the review corpus become the features to the classification model.

3. **Bag of words model**

In the bag of words or the unigram model, the review texts are converted to term frequency vectors, where each dimension represents the count of the term in the review text. The term frequency representations of the entire corpus become the features to the classification model.

4. **Bigram model**

In the Bigram model, the review text is broken down into bigrams - set of two words occurring consecutively in the text and the document is then represented as the term frequency vector of these bigrams. Each dimension of the term frequency vector represents the count of the bigram in the review text.

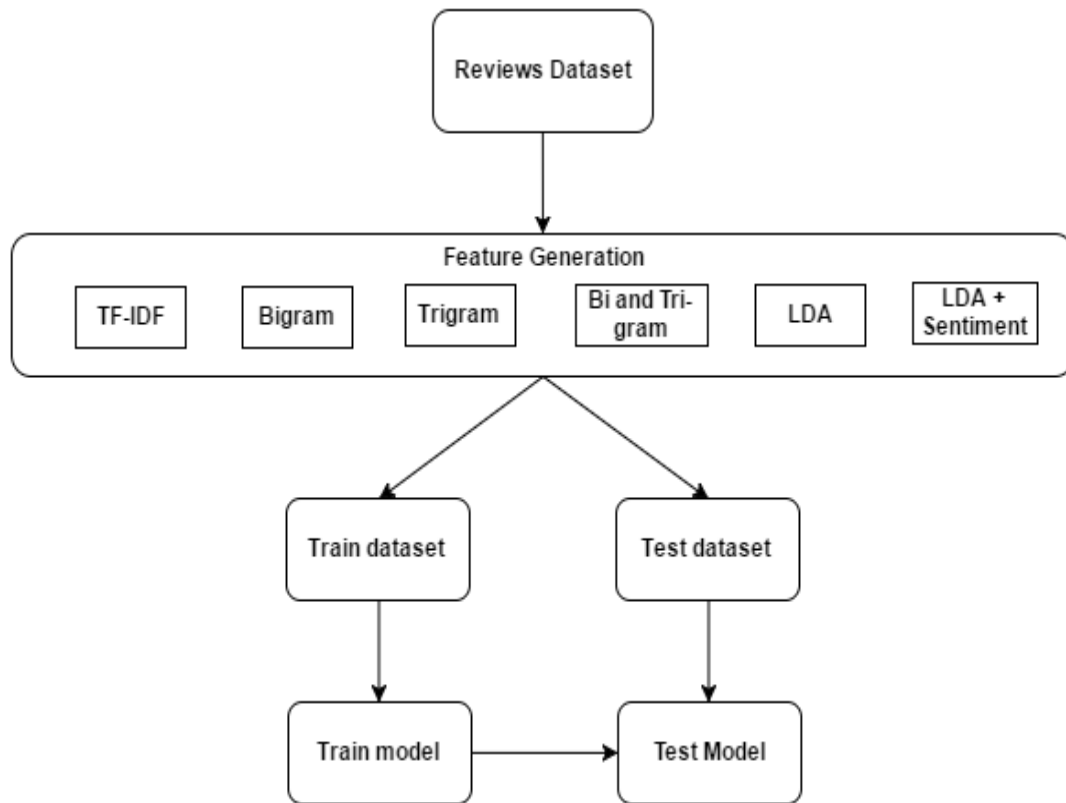


Figure 4.1: Flow chart showing the models in the workflow

5. Trigram model

In the Trigram model, the review text is broken down into trigrams - set of three words occurring consecutively. The term frequency vectors of these sets become the features of the review text where each dimension represents the count of the trigram in the review text.

6. Bi and Tri-gram model

In the Bi and Tri-gram mode, the review text is broken down into both bi and tri-grams. The term frequency vectors of these sets become the features of a review text where each dimension represents the count of the bigram or trigram in the review text.

7. LDA model

All the previous models consider all the words/bigrams/trigrams and their frequency as the training features. In the case where the corpus is huge, the dimensionality of the input features would be huge. To reduce the dimensions, we have used the Latent Dirichlet allocation (LDA) algorithm which discovers a given number of topics or themes in the corpus. Each review text is then represented as vector of probability distribution across these topics. These topic distribution vectors become the features to the classification model. We have used the `gensim` package in python to create the LDA model:

```

1 {
2     ldamodel = gensim.models.ldamodel.LdaModel(corpus,
3         num_topics=15, id2word = dictionary)
  }
```

Parameters:

- `num_topics`: The number of topics that should be discovered.
- `id2word`: The dictionary created from the corpus

8. **LDA model + Sentiment Layer** In this model, sentiment of the review was fed in as the feature alongside the topic distribution of the review. The sentiment of the review was extracted using the Naive Bayes classifier.

5 MODEL EVALUATION

Table 5.1: Baseline model performance metrics

Accuracy	0.273578
F1 score	0.117535
Precision	0.074845
Recall	0.273578

We can see in the above table that the Baseline model, which predicts every review's rating to be the average rating of all reviews, performs poorly in terms of accuracy. Even the precision of the model is only 7.5%.

Table 5.2: TFIDF model performance metrics

Metric	Logistic Regression	Multinomial Naive Bayes
Accuracy	0.640334	0.553686
F1 score	0.640334	0.553686
Precision	0.618353	0.507628
Recall	0.640334	0.553686

In the TF-IDF model, we tried two classification models - Logistic Regression and Multinomial Naive Bayes. From the above results, we can see that Logistic Regression outperforms Multinomial Naive Bayes in terms of all metrics. The best accuracy obtained using this model is 64%.

Table 5.3: Bag of words model performance metrics

Accuracy	0.630097
F1 score	0.624512
Precision	0.621452
Recall	0.630097

Table 5.4: Bigram model performance metrics

Accuracy	0.63593
F1 score	0.61031
Precision	0.61305
Recall	0.63593

The bag of words, bigram, trigram and bi & tri-gram models were built using Multinomial Naive Bayes classifier and comparing their performance from Table 5.3, 5.4, 5.5 and 5.6, we can see that bigram model has the best accuracy out of the 4 models, which is at 63.59 %.

Table 5.5: Trigram model performance metrics

Accuracy	0.611375
F1 score	0.573271
Precision	0.581688
Recall	0.611375

Table 5.6: Bi and Tri-gram model performance metrics

Accuracy	0.614550
F1 score	0.573366
Precision	0.584915
Recall	0.614550

Table 5.7: LDA model performance metrics

Metric	AdaBoost	Logistic Regression	Multinomial Naive Bayes	Random Forest
Accuracy	0.486046	0.482509	0.447089	0.468679
F1 score	0.418104	0.408028	0.276263	0.420345
Precision	0.42255	0.399447	0.199888	0.409459
Recall	0.486046	0.482509	0.447089	0.468679

Table 5.8: LDA + Sentiment model performance metrics

Metric	AdaBoost	Logistic Regression	Multinomial Naive Bayes	Random Forest
Accuracy	0.556981	0.556569	0.447352	0.544579
F1 score	0.51275	0.500961	0.276845	0.513781
Precision	0.519836	0.516783	0.281885	0.511234
Recall	0.486046	0.482509	0.447089	0.407346

The LDA model didn't perform as well as our expectation to predict the rating from the review text. From Table 5.7 we can see that using the topics extracted using LDA as features only leads to a best accuracy of 48.6%. We attributed this poor performance to the fact that the topics extracted do not have any sentiment associated with them. So, in our next model, we added a sentiment layer as an additional feature to the model. The sentiment of the reviews was extracted using the Naive Bayes classifier. As the Table 5.8 indicates, the LDA + Sentiment model performs much better than the LDA model and we achieve a best accuracy of 55.6%.

Multiple classifiers were tested for the LDA model, such as - Adaboost, Logistic Regression, Multinomial Naive Bayes and Random Forest. We observed that the Adaboost classifier gives the best accuracy and it is closely followed by Logistic Regression classifier.

6 CODING CONTRIBUTION

6.1 DATA TRANSFORMATION PRE-PROCESSING

Since the files were in JSON format, they were first converted to CSV format. After converting to CSV format, the business and review data frames were joined on *business_id* for restaurant establishments. From the combined dataset, we got over 1.6M reviews for restaurants alone. The dataset now consisted of just two columns : *Review Text & Star Rating*.

```
1 {  
2     len(resto_review_data)  
3     1630712  
4 }
```

The length of reviews was tweaked based on our observations from Exploratory Dataset Analysis. Since a very small review might not contribute meaningfully to our models, we decided to limit the minimum characters in a review to 50, while the maximum character limit was fixed to 500 (this was kept low, because at higher ranges, the dataset was still too huge and required more computing power to run than was possible on our MacBooks). This further reduced the number of records to a little over 900k.

```
1 {  
2     resto_review = reduceReviewBasedOnLength(resto_review_data=  
3         resto_review_data, minReviewLen=50, maxReviewLen=500)  
4     len(resto_review)  
5     910340  
6 }
```

6.2 CLEANING

Since we are dealing with real world text reviews, the data available to us will contain plenty of punctuation words as well common English stopwords. We followed a standard process for text mining to further cleanse and process the review dataset. In this, we first converted all the reviews to **lowercase**. We then removed **numbers & punctuations**. After this, we **tokenized** the text using the NLTK library's tokenizer. Post this, we removed the **stopwords** and we also performed **stemming** of the words. The cleaned dataset was thus used as an input to all our models trained.

6.3 CREATING TRAINING AND TESTING CORPUS

We used **70-30** sampling to create a training and testing corpus from the cleaned text reviews. In order to ensure sufficient representation of all the star label values in our training corpus, we split the **original corpus on the basis of the star labels**, thus creating 5 corpora, one for each star label. The 70-30 sampling was then carried out on each of the 5 corpora and the training and test files thus created were then combined to create one big training and testing corpus. The number of reviews in training corpus for each star label is given in Table 6.1.

6.4 LDA MODEL DEVELOPMENT

The LDA[4] model is present in **gensim package** in python. The inbuilt library method was not so straightforward and required a **vectorized bag of words** corpus as an input. It also required a dictionary developed from the available training corpora. The parameters that could be tweaked while developing the model were the corpora size and the total number of topics we want to extract. We chose the **total topics as 15**. In a normal vectorized corpus, the dimensionality would have

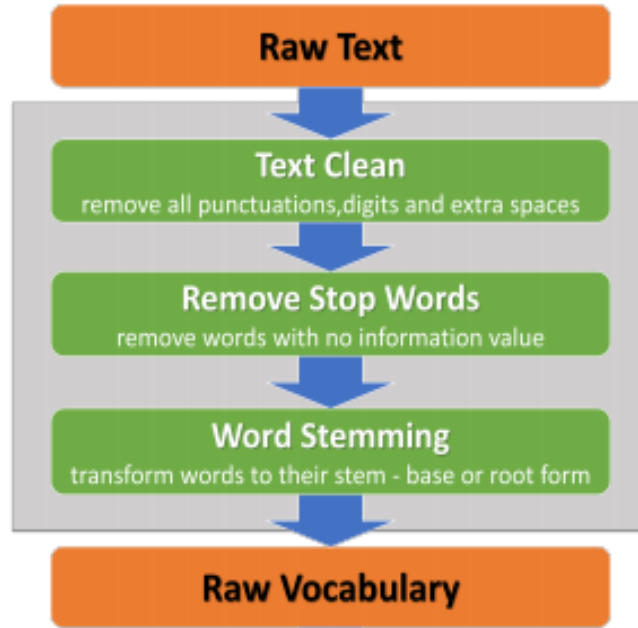


Figure 6.1: Data Cleanup Flow

Table 6.1: Star rating distribution in the training corpora

Dataset	Count
Rating 1	58739
Rating 2	47154
Rating 3	72107
Rating 4	174334
Rating 5	284902

been the entire size of the disctionary, which is very huge. Selecting the total topics essentially will reduce the dimensionality of our training corpora to merely 15 selected topics. The topic probability distribution dataset was used as a feature to create new training corpora which was used to train off the shelf classifiers such as *MultinomialNaiveBayes*, *LogisticRegression*, *RandomForestClassifier*, *AdaBoostClassifier*. The performance of the models is discussed in a separate section on Model Evaluation.

6.5 CONTRIBUTIONS FROM TEAM MEMBERS

All team members had equal contributions to this assignment. (33% each). The key contributions are as follows :

1. Karan Grover.
 - Performed the conversion of json to csv.
 - Suggested the merger of business and reviews file into one.
 - Performed the cleaning and pre-processing of the dataset.
2. Pragya Arora
 - Performed the exploratory analysis of the dataset, and suggested using a limit size range of reviews.

- Suggested and worked on Bi-Grams, Tri Grams & Bag of Words approach.
- Jointly contributed with Piyush in TF-IDF Vectorized model.

3. Piyush Ghai

- Suggested to keep the representation of all star labels same in the training and testing split.
- Suggested and worked on LDA & LDA + Sentiment Analysis approaches for classification.

7 TOOLS AND TECHNOLOGY STACKS

For this assignment, we've used a variety of tools. The coding for this assignment was done using Python. The following is an exhaustive list of modules/tools used :

- Python 2.7
- Pandas (For manipulating the dataframes)
- nltk (For corpus, dictionary, stopwords, stemming etc)
- NumPy (For classification model results)
- seaborn (For graphs)
- cPickle (For saving pickle files and using them later in models)
- iPython (The interactive python shell, which was used for development)
- sci-kit learn (For various classification algorithms)
- Gensim (For LDA model)

REFERENCES

- [1] Yelp is an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations and online food-delivery service Eat24. <https://en.wikipedia.org/wiki/Yelp>.
- [2] The Yelp Dataset Challenge provides the academic community with a real-world dataset over which to apply their research. More about yelp dataset can be found on : https://www.yelp.com/dataset_challenge.
- [3] NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries. <http://www.nltk.org/>.
- [4] In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation.