INDIANA UNIVERSITY
Social Media Mining
ILS Z639
Restaurant Categorization & Star Rating Prediction Using Yelp Data

Deepika Bajpai (dbajpai@umail.iu.edu), Hemant Pandey (pandeyh@umail.iu.edu)
School of Informatics and Computing

Abstract - In this study, we analyze the Yelp dataset with an objective of determining the stance of a particular review by a user with regard to the positive, negative or neutral nature of the review towards a particular restaurant. Further analysis includes star rating prediction and finding interesting trends between various restaurants under study. This report is also a compilation of various stance detection procedures computed over Yelp reviews using computational linguistic analysis tools & techniques. This study is relevant because of its potential to accelerate or revamp the business strategy of a particular restaurant based on the yelp reviews it received.

I. INTRODUCTION

It is usually believed that in order to judge something or someone and take actions involving the thing or person under study, one should consult facts or information about the thing or person. These facts or information can also be available in the form of reviews or opinions generated by other human beings. Human Beings hence, prefer making choices, opinions or perceptions based on the opinions generated by other individuals. Although many such reviews can be faulty yet they do play an impact in making other individual's mindsets regarding particular thing or person.

In a world where social media reviews greatly influence a person before he/she considers purchasing a good or service, Yelp is one such application that caters to this need for the food industry. Yelp is an American corporation that generates crowd sourced reviews about local businesses on the basis of various factors that can be used to characterize the business. The yelp application allows users to read/write reviews about their personal experiences with a particular restaurant that they had visited. Based on these mass reviews, a person can decide whether he/she should consider a particular restaurant or not. The Yelp Data set is one of the most interesting data set as it allows one to find many interesting insights about the eating patterns of individuals towards food belonging to different parts of the world.

II. PROBLEM DESCRIPTION

This project caters to some problems involving sentiment analysis as well as classification based on certain probabilistic models. Some of the problems that our projects deals with are as follows:

A. Prediction of Star Rating Using Review Text

The Yelp application requires it's users to provide Star Ratings to a particular restaurant on the basis of the experience they had with a particular restaurant during a particular visit. The star rating of a restaurant is generally an indication of whether the restaurant is good or bad as per our analysis. In this project, we aim at predicting the star ratings given the review by using the classification techniques of support vector machines with linear kernel and Naive Baye's.

B. Categorization of restaurants as good, bad or average

Reviewers usually feel very burdened when they need to glance through long chunks of texts, at times consisting of irrelevant information just to infer whether is it worth spending time and money at a particular restaurant or not.
This project aims at categorizing restaurants as good, bad or average by performing sentiment analysis on the review text and identifying whether a given review is positive, negative or average, eventually associating it with a particular business (restaurant) via the business_id field from the business data set. A particular review is categorized as good bad or average depending on the number of occurrences of positive, negative and neutral words in the review text. We are expecting that reviews which have a higher star ratings will turn out to be positive and hence will be under good category and reviews with lower star ratings will be negative and hence under bad category.
To predict the category, we are tagging the reviews with actual rating as 4 or 5 to be good, 1 or 2 as bad and 3 as average.We are using the the Tf-idf vectorizer for feature selection considering each review text and to give less weightage to stop words.

C. Insights or Interesting trends among different restaurants

Based on the various attributes offered by a restaurant to its customers, one can analyze many interesting trends or insights about a particular restaurant as well as the people visiting the restaurant. For instance, a restaurant that serves alcoholic beverages has more customers than a restaurant that does not serves alcoholic beverages. One can also find insights between restaurants serving different types of food, like American vs. Mexican. Such information is very helpful

for study as well as the business itself as it would allow the business owners to modify their strategy accordingly. Future diners can be made aware of a restaurants service, specialty, etc. Selection and recommendation of restaurants to individuals using the Yelp application can also be improved on the basis of these insights.

III. RELATED WORK

Many past projects were consulted for inspiration required to complete this project successfully. In order to determine the stance of a review, we need to perform sentiment analysis over the data set. One such work that inspired us was by Paul Ekman(1992)[2], according to whom the 6 types of emotions required for emotion mining of a stance are Anger, Sadness, Disgust, Fear, Enjoyment, Surprise. Many stance detection techniques like the once mentioned by Somasundaran & Wiebe(2010)[3] or (M. Mohammad et al., n.d.)[4] over twitter tweets were being used to find the stance of the review text.

For star rating prediction we are using the Naive Bayes classifier (Rish, n.d.)[5] and Support Vector Machines. We are implementing these two classifiers using the libraries present in Scikit-learn package for Python (Scikit-learn.org, n.d.)[6].

Another work that we referred to is for Multiclass sentiment analysis with restaurant reviews(Moontae lee,Patrcik Grafe)[9] which makes use of unigrams,bigrams and trigrams and classification techniques using sentiments to predict the ratings from one to five stars.

Mining Millions of Reviews: A Technique to Rank Products Based on Importance of Reviews[Kunpeng Zhang, Yu Cheng,Wei-keng Liao,Alok Choudhary][10] In this paper, the model applies weights to reviews in order to generate a product ranking score, making it more feasible to mine large review data sets.

Another such work is by Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowl-edge Discovery and Data Mining, KDD 04, pages 168177, New York,NY, USA. ACM[11] In this work, the authors aims to mine and summarize all the customer reviews of a product.and provided different techniques to perform three essential tasks.: (1) mining product features that have been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results.

IV. DATA SET

There are 2 Data Sets under consideration, namely: the Business Data set and the Review Data set.

A. Review DataSet

The Review Data consists up of the various reviews given by a particular user to a particular restaurant at a particular visit star rating provided by visitors to the restaurants. The features include business_id, rating, review, user_id, review_id, etc.

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

Fig. 1. Review data-set in JSON format

B. Business DataSet

The Business Data consists up of various attributes required to characterize a restaurant and its service. For Example, acceptance of credit cards, liquor license, ambiance, etc. The business data set also consists of the overall Star Ratings of the Restaurant. It contains the business_id, category, name, star ratings, etc.

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

Fig. 2. Business data-set in JSON format

C. Data Merging

Since both the data sets consists up of a common field business_id acting as a primary key between them, we hence merged both the data sets into a single data set for easy manipulation. The first step in data merging includes the conversion of JSON data sets to CSV using python code provided by yelp using the 'utf-8' encoding.The next step was to read the data sets as pandas data frames and then performing an inner join on business_id field and converting the resulting data set to a csv file. Once we had our merged csv file ready, we selected the data pertaining to business_ids having the categories as "Restaurants" only and then further splitted our data into Mexican and American

restaurants categories. In the end we were left with two data sets one for each Mexican restaurant category and another for American restaurant category.
The next step was to label the data. For this we appended a label value as -1 (bad),0 (average) and 1 (good) for three scenarios where star rating is either 1 or 2 as first scenario, star rating is 3 as second scenario and star rating is either 4 or 5 as third scenario respectively.

D. Data Pruning

The original Mexican and American data set was highly skewed as most of the samples belong to rating 4 and 5. It can be depicted in figure 3 and 4. Also Since we have more

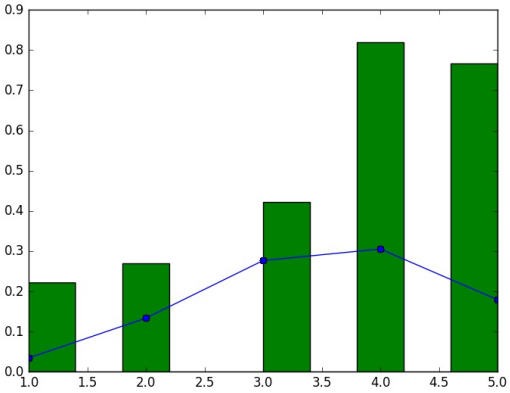

Fig. 3. Histogram of Star rating distribution in American restaurants

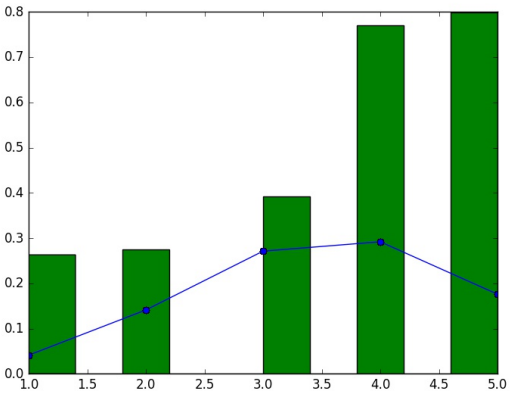

Fig. 4. Histogram of Star rating distribution in Mexican restaurants

than two million reviews in American restaurant category and around a million reviews in Mexican restaurant category, extracting features from all of them and building a classifier with that amount of samples it is computational expensive and, in some cases, even impossible. Because of this, we extracted a reduced amount of reviews for each of the restaurant category by selecting a data set that has around 10,000 representative samples from all three categories totalling upto around 30,000 data rows.
Figure 5 depicts the distribution of star ratings in the American restaurants in our pruned dataset. Similarly Figure 6 shows the review sentiment class distribution where we have equal reviews belonging to each of the three classes i.e -1 ,0 and 1. Figure 7 and 8 presents the star rating and sentiment

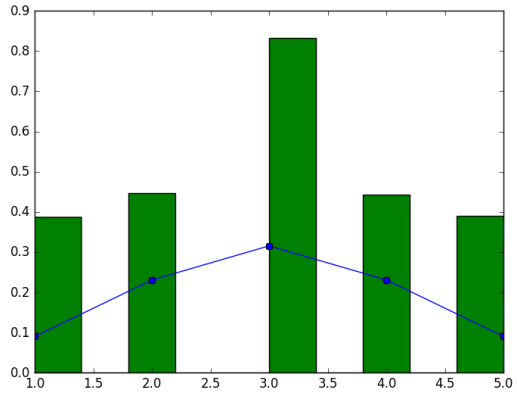

Fig. 5. Histogram of Distribution of star rating in American restaurants(pruned data set)

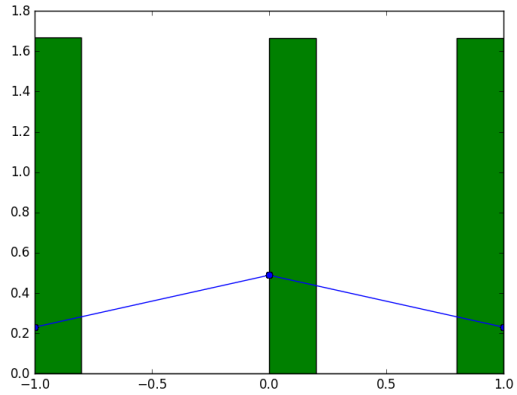

Fig. 6. Histogram of Distribution of review sentiment in American restaurants(pruned data set)

distribution representation for pruned mexican restaurants dataset.

V. METHOD & DATA SPLITTING

A. Stop Words

Stop words are the most common words of the review text and hence are given less weightage. Articles such as a and the and pronouns such as he, and I etc. do not provide much information about sentiment. The basic idea is that if a word occurs so many times in one documents, its important but at the same time, if it occurs in so other documents also so many times, it becomes less important and should be considered as a stop word. By removing these

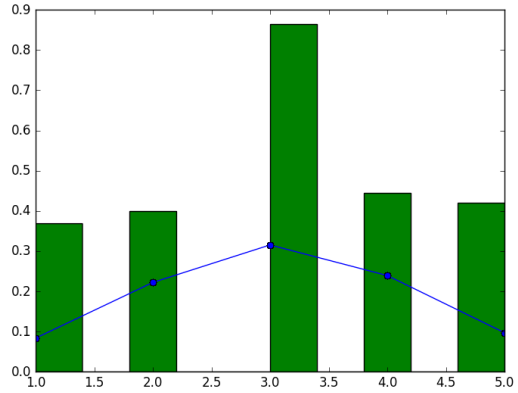

Fig. 7. Histogram of Distribution of star rating in Mexican restaurants(pruned data set)

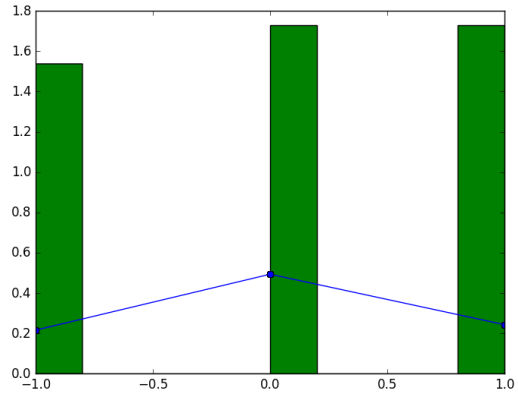

Fig. 8. Histogram of Data distribution of review sentiment in Mexican restaurants(pruned data set)

stop words, we can better focus on the more useful features .These can be removed by calculating Tfidf. Tfidf ,short for term frequencyinverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. We have used the Tf-IDF vectorizer from sklearn package which converts a collection of raw documents to a matrix of TF-IDF features. The packages used are[12] :

- CountVectorizer which Tokenize the documents and count the occurrences of token and return them as a sparse matrix.
- fit_transform which learns the vocabulary dictionary and return term-document matrix.
- TfidfTransformer which applies the Term Frequency Inverse Document Frequency normalization to a sparse matrix of occurrence counts.

So after performing this step, we got the features for applying our classification models.

B. Classification techniques

The task of predicting the star ratings and then categorizing as positive, negative or average is done via two linear classifiers i.e Multinomial Nave Bayes (MNB) and a Support Vector Machine (SVM) classifier using the Scikit learn python package. We trained both classifiers with 80% of the data and tested them with the other 20% of the data to calculate the accuracy.

C. Multinomial Naive Bayes (MNB):

Often used in Machine Learning, Naive Bayes is a probabilistic classifier that applies Bayes Theorem of Probability with strong (naive) independent assumptions between the features. In other words it is used for text classification of training data based on the classification of test data. When our data set under study is multinomially distributed, it is known as multinomial distribution. In such a distribution, the distribution is parameterized by vectors. In this project Naive Bayes classifier is being used as one of the classifiers to perform the star rating prediction. It is a simple classification method based on the Bayes rule. We have the set of feature vectors obtained using the Tf-idf vectorizer in the previous set and the class labels as -1,0 and 1 which corresponds to bad,average,and good. The model will learn the pattern on the labeled training data and predict for test data.

D. Support Vector Machine (SVM):

Another classification algorithm that we used for the prediction of star ratings in the Yelp data set is Support Vector Machines or SVM. Support Vector Machine is a type of Supervised Learning Technique. The algorithm performs classification by creating a set of hyper-planes between the various data points which need to be classified. These hyper-planes allows us to visualize SVM models as points in the coordinate system forming clusters separated from each other via distinct gaps created by the hyper-plane itself.The SVM classifier helps to use the existing dataset as a training model to predict values on the test data.The selected features from a review will help to accurately divide the data into the category as being good,bad or average.
In our project we have used linear kernel SVM. Due to the high dimensionality of vector space, the kernel trick is rendered unnecessary and hence for bag of words model data is generally linearly separable with acceptable accuracy. We have done this by implementing the classifier using Linear SVC from scikit-learn python library. It creates a set of classifier per pair of class and then these classifiers vote to find the most confident class.In our case, we have three classes.

E. Cross Validation:

Also known as Rotation Estimation, a K-fold Cross Validation technique splits the training data set into k-smaller sets. For each of these k-folds the following procedure is followed:

- A model is trained using K-1 of the folds as training data.

- The resulting model is then verified on the remaining part of the data.

We are using simple cross validation by converting or total data set to two parts. The 80% of the labeled data set is used for training purpose and the remaining 20% we are using for testing the accuracy.

VI. EVALUATION

The accuracy was calculated using the classification report from sklearn.metrics package of python.The different parameters through which the accuracy has been measured is:

A. Precision:

The precision is calculated as the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is basically the ability of the classifier not to label as positive a sample that is negative.

B. Recall:

The recall is the calculated as the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is the ability of the classifier to find all the positive samples.

C. f1-Score:

The f1-score gives us the harmonic mean of precision and recall and hence is kind of a balance between the two.

D. Support:

The support is the number of samples of the true response that lie in that class.

E. Results Obtained:

The figure 9,10,11,12 presents the different accuracy results obtained for Mexican and American data sets using Naive Byes and SVM classifiers.We can see that SVM classifier has performed slightly better than the Naive Bayes alsthough both of them are used for linear classification here.It could be attributed to the large dimension of the problem at hand.

class	precision	recall	f1-score	support
-1	0.82	0.64	0.72	1797
0	0.58	0.77	0.66	1999
1	0.82	0.72	0.77	1981
avg/total	0.74	0.71	0.72	5777
The accuracy score is 71.09%				

Fig. 9. Accuracy using Naive Bayes for Mexican restaurants

class	precision	recall	f1-score	support
-1	0.73	0.76	0.75	2012
0	0.61	0.66	0.63	2008
1	0.84	0.73	0.78	1980
avg/total	0.74	0.71	0.72	5777
The accuracy score is 71.73%				

Fig. 10. Accuracy using Naive Bayes for American restaurants

class	precision	recall	f1-score	support
-1	0.81	0.78	0.79	1797
0	0.68	0.73	0.7	1999
1	0.84	0.81	0.83	1981
avg/total	0.78	0.77	0.77	5777
The accuracy score is 77.24%				

Fig. 11. Accuracy using SVM with linear kernel for Mexican restaurants

class	precision	recall	f1-score	support
-1	0.8	0.8	0.8	2012
0	0.66	0.69	0.67	2008
1	0.83	0.8	0.81	1980
avg/total	0.76	0.76	0.76	6000
The accuracy score is 76.07%				

Fig. 12. Accuracy using SVM with linear kernel for American restaurants

VII. DISCUSSION & CONCLUSIONS

The project helped us in observing few interesting trends.The first observation is regarding the accuracy of the classifiers. SVM performed slightly better than the naive Bayes classifier in terms of accuracy. In terms of comparison between American and Mexican restaurants, the prediction accuracy is almost similar. This could be attributed to the reason that we have selected similar sample distribution in both of them.

A. Mexican vs American Restaurants

One interesting insight that we noticed was that Mexican restaurants belonging to states much closer to Mexico (West Side) or having a larger number of Mexican population in them, received a higher average star rating for Mexican restaurants in that state than restaurants belonging to states farther from Mexico (East Side) or having a lesser number of Mexican population in them. This experiment was conducted on 28881 Yelp data records for Mexican Restaurants and from its results we can conclude that Regions much closer to Mexico have average star ratings starting from 3.47 and ranging to 3.9259. Whereas, for regions farther from Mexico it starts from 3.1309 and ranges to 3.51119. Thus, one can say that regions near to Mexico have a better understanding of the culinary dishes than the ones that are far off. This can be corroborated from the statistics mentioned in figure 13.

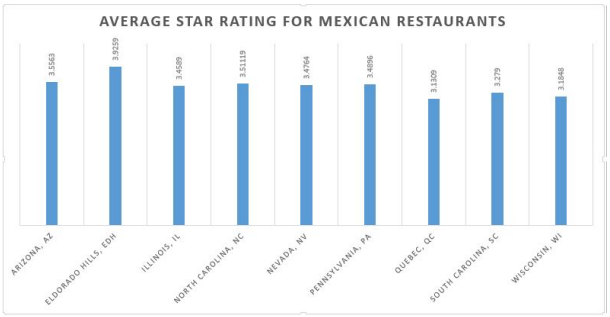

Fig. 13. Average star ratings across various regions for Mexican food

B. Alcohol Serving vs Non-Alcohol Serving Restaurants

We calculated the average star ratings for restaurants serving alcohol and the average star ratings for restaurants not serving alcohol. The difference between the two ratings is shown in Figure 14 Form the plot given in Figure 14,

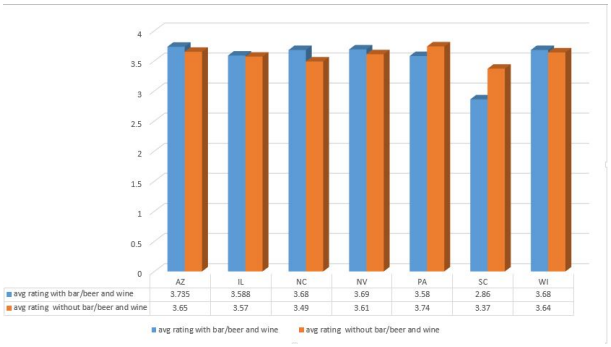

Fig. 14. Alcohol Serving vs Non-Alcohol Serving Restaurants Average Ratings Plot

we can say that the average rating of restaurants serving alcohol is slightly higher than the restaurants not serving alcohol. This assumption however is proved wrong for 2 states (South Carolina & Pennsylvania), which we are considering to be outliers. This observation of ours can be corroborated from the fact that since acquiring liquor license is an expensive deal and only restaurants with good standards are able to afford it, therefore they would try their best to provide the best experience possible to the customers, hence the better ratings. Whereas, the restaurants not serving alcohol could include fast food joints as well as small cafeterias.

In this project,we can conclude that the star rating of the restaurants can be predicted via performing the sentiment analysis on the review text and by applying the classification techniques. They can further be classified into different categories like good , bad or average considering the sentiments of the review text.

VIII. Future Work

The model for categorizing the restaurants into good,bad or average category proposed in the paper successfully manages to classify the reviews. But still there is lot of scope for improvement. The results can be further improved by considering bigrams, trigrams, quadgrams in the review text. Also we can consider the helpfulness of reviews as provided in the data set as another feature for a better prediction. Another enhancement that we can further is to detect the fake reviews and apply our prediction model for better results. Apart from the interesting trends/insights mentioned by us, there is scope for finding many other insights that would greatly characterize the eating patterns of people belonging to a particular region with respect to foods from some other part of the world.

ACKNOWLEDGMENT

The completion of this project was only possible due to the constant guidance and support of Professor Allen Riddell and our classmates. We would also like to thank your fellow classmates who inspired us by sharing their ideas during group presentations. There are many books and publications that inspired us towards the completion of this project. References of these books and publications has been mentioned in the references section.

References

[1] Jurafsky, D. and Martin, J. (2009). Speech and language processing. 1st ed. Upper Saddle River, N.J.: Pearson Prentice Hall.
[2] An Argument For Basic Emotions. 1992. Computing Reviews, 24(11):503512.
[3] Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 116124.
[4] M. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. and Cherry, C. (n.d.). SemEval-2016 Task 6: Detecting Stance in Tweets.
[5] Rish, I. (n.d.). An empirical study of the naive Bayes classifier.
[6] Scikit-learn.org. (n.d.). 1.4. Support Vector Machines scikit-learn 0.18.1 documentation. [online] Available at: http://scikit-learn.org/stable/modules/svm.html.
[7] Scikit-learn.org. (n.d.). 1.9. Naive Bayes scikit-learn 0.18.1 documentation. [online] Available at: http://scikit-learn.org/stable/modules/naive_bayes.html.
[8] Scikit-learn.org. (n.d.). 3.1. Cross-validation: evaluating estimator performance scikit-learn 0.18.1 documentation. [online] Available at: http://scikit-learn.org/stable/modules/cross_validation.html.
[9] Tay, R. (2016). Unigram-based Sentiment Analysis of Textual Restaurant Reviews for Prediction of Above Average Rating. [online] Rstudio-pubs-static.s3.amazonaws.com. Available at: https://rstudio-pubs-static.s3.amazonaws.com/127807_4a271a7d9a4343f89595dae16ea1988e.html.
[10] En.wikipedia.org. (2016). Sentiment analysis. [online] Available at: https://en.wikipedia.org/wiki/Sentiment_analysis.
[11] Available at:http://dl.acm.org/citation.cfm?id=1014073