DALHOUSIE UNIVERSITY
Faculty of Computer Science

CSCI 6509
Natural Language Processing

Project Report

April 6th, 2020

P-25: Analysing Yelp Reviews using NLP Techniques

Name	Email
Haard Prashant Shah	hr606438@dal.ca
Nishant Amoli	ns573209@dal.ca

Abstract

In this project, Yelp restaurant reviews were analysed using NLP techniques. Out of 6.6 million reviews, we selected only Ontario restaurants with more than 300 reviews. We divided the 1-2-star reviews as negative reviews, 3-star reviews as average review and 4-5-star reviews as positive reviews. After removing stop words and performing TF-IDF, various topic modelling approaches were used to find out relative positive, average and negative weight of topics for different restaurants. The results obtained were visualized using Cognos BI [1]. Moreover, we also compared the running time complexity to for topic modelling using Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

1. Introduction

1.1 Significance of project

If 100 restaurants were launched at the beginning of the year, then only 40 of them would survive at the end of year and only 20 out of all 100 would be able to reach their 5th year anniversary as shown in figure 1 [2]. The primary reason for this closure is the lack of self-awareness of the restaurant owners [3]. The owners do not know what they are doing right and what they are doing wrong. What if the owners can know about what to improve in the restaurant in order to increase the business and listen to their customers? One possible way to solve this problem is to read the reviews of the restaurant. But what if there are thousands of reviews for the restaurant? It becomes tedious and almost impossible to study all the reviews and find some similarity between to identify the topic. This is where NLP comes to the rescue. Topic modelling is used to automatically identify the hidden topics in a document (review) and derive the patterns present in the corpus of text. Using topic modelling techniques of NLP, a business owner can know what are the most talked things about their restaurant in positive reviews, average reviews and negative reviews.

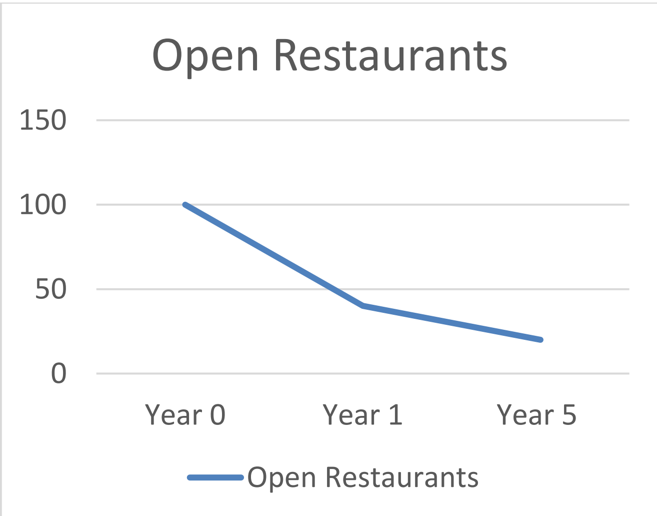


Figure 1: Number of restaurants surviving initial period

1.2 Project Objectives

a. To extract relevant reviews for analysis.
b. Perform TF-IDF on the reviews of selected restaurants.
c. To perform topic modelling of reviews using LDA and analyse the result.
d. To perform topic modelling of reviews using NMF and get appropriate topics for different categories of the reviews.
e. Analyse the time taken to perform topic modelling using different techniques with same hardware resources.
f. Visualize the results of topic modelling for different categories of reviews and for different restaurants.

2. Related work

Significant amount of work has been done regarding topic modelling and analysis of Yelp dataset. Such research includes sentiment analysis on dataset, predicting restaurant rating based on collection of reviews. The primary focus for our project is not predicting ratings or performing sentiment analysis but to find the reason behind a restaurant's success and failure. Some of the research papers that would help in building the motivation for this project is as follows.

a. A research paper on "How to Turn Customers Into Advocates" [4] shows the effect of positive Yelp reviews on overall business of the restaurant. When analysed over a period, it was found that when the rating of a restaurant increased by 1 star, the business revenue had a huge increase of 5-9%. Thus, we can say that if a business with a turnover of $1 million, could increase the rating by 2 stars, it would lead to significant increase in the yearly profit. This analysis forms our basis of motivation for topic modelling of Yelp reviews. If the business owners can recognize the value of Yelp reviews and can quickly get an update about the good and bad reviews of restaurants, they can save the business from closing.
b. Topic modelling has number of approaches to it. Among them, an older and more basic approach is Latent Semantic Indexing (LSI). This method benefits from semantic structure and uses it to efficiently retrieve essential document topics. The paper "Indexing by Latent Semantic Analysis" [5] performs LSI in an efficient way. However, in case of SVD, if two-dimensional plot of terms and documents containing those terms was constructed by applying the LSI model, it was found to produce not so great results. Even the documents which that had no common terms were near to each other. However, the implementation of LSI is budget friendly as each document and term needs to be represented by order of 50-150. However, it does not provide functionality to map reduce like other methods. It is very slow on large corpora and accuracy could be compromised. There could be a trade-off with amount of resources by applying SVM and other supporting techniques.

c. Another topic modelling technique called Latent Dirichlet Allocation (LDA) was used to extract subtopics from Yelp reviews in the paper "Improving Restaurants by Extracting Subtopics from Yelp Reviews" [6]. This paper performs analysis of Yelp reviews of restaurants using LDA as primary method. By using LDA, hidden star ratings of restaurants can be predicted. Without dividing the reviews into positive and negative, if LDA is applied on the collection of reviews, it results in most frequent sub-topics like “service”, “value”, “take out”. Hence restaurants with a given star rating may have higher or lower hidden star rating for subtopic like service. As this is unsupervised learning, determining its accuracy manually is again a tedious work. Moreover, if the reviews had been divided, a better understanding of topics polarity could have been gained. Moreover, LDA performed iteratively for 1-star and 5-star review leaves out lot of reviews as overall clustering of reviews is more for 3-4 stars. Hence there were some factors which were not considered in this research and thus, provides us a with a good room for improvement.
d. Third approach to topic modelling is Non-Negative Matrix Factorization (NMF). This is comparatively newer approach and has been found more efficient in terms of interpretation of results. According to study about “Document Clustering Based on Non-negative matrix Factorization” [7], NMF assumes that a document consist of K-clusters and one topic heavily dominates a document or several topics contribute towards a document. NMF finds the positive factorization of a given vector and hence results are not distributed in negative space. The research implemented two spectral clustering methods namely Average Association and Normalized Cut to determine the improvement in accuracy. This is better approach then previous manual analysis in LDA. These methods model the documents using undirected graph. The analysis found a high accuracy for clustering documents. Hence, document clustering results can be directly obtained without performing any more clustering options. Thus, we decided to use NMF as our primary method for topic modelling the Yelp reviews. This research was a proof that even in worst performance scenario, NMF would still give comparable and interpretable results.
e. Another research that focussed on analysing Yelp reviews was “Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews” [8]. Using SVM, accuracy of 88.9% was obtained while using TF-IDF resulted in accuracy of 88%. The reviews were categorized but the results were not of much interest to business owners. As expected with opinion mining, words like friendly were top-rated in most of the restaurants. The accuracy can still be improved using neural networks and random forest but that would still defy our aim of finding topics of interest which can help in improving the restaurant. Also, for such approaches and big data, more hardware resources are required for better efficiency. Thus, topic modelling using NMF and LDA seems a much more relevant and feasible approach that could possibly help to achieve our project goal.

3. Implementation

3.1 Defining the problem precisely

The first step gaining proper results is to know exactly what is required to be done as a part of solution. Hence, the problem statement can be defined as performing topic modelling on selected restaurant reviews so that we have weighted topics for each restaurant among positive, average and negative reviews.

3.2 Dataset

Yelp has a huge json collection of reviews of various businesses. To be specific, the total review size is about 6.5 GB which includes reviews of hotels, restaurants, parks and various other shops. For the purpose of this project, we selected only restaurant reviews. After filtering the reviews to only restaurants, there were still 6.6 million reviews [9] for provinces like Arizona, Ontario. The following pie chart shows the distribution of reviews for various provinces.

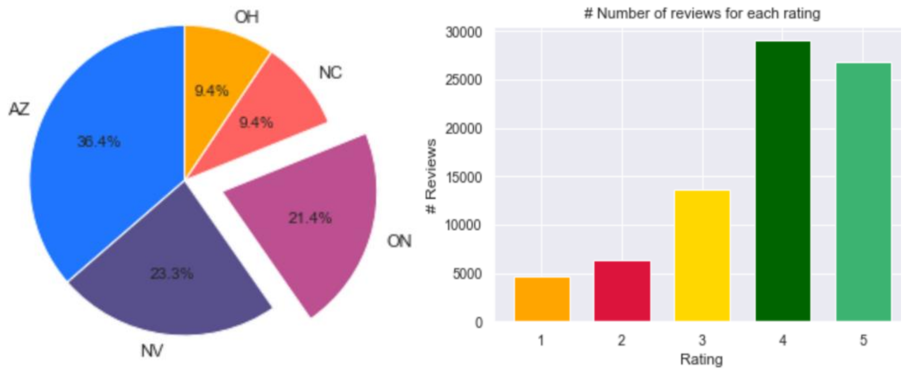


Figure 2: Review distribution

We can see that Ontario has ideal number of reviews for analysis (considering the hardware constraint). Hence, Ontario restaurants were selected for topic modelling the reviews. However, there were restaurants with review count of 20 as well as those with the review count of 1500. Topic modelling the restaurants with lesser review count would lead to inappropriate results. Thus, restaurants with more than 300 reviews were selected. More than 170 restaurants had greater than 300 reviews for further analysis.

3.3 Pre-processing

3.3.1 Cleaning and Stop Words Removal

The reviews were cleaned to remove special characters, URL's and emojis. For topic modelling and TF-IDF, it is essential to remove stop words which may hinder our results. Using standard sklearn library, stop words were removed. However, removing these standard stop words was insufficient as the reviews were heavily loaded with words like "amazing", "know", "pretty". These words may be of some value for sentiment analysis, but they are of no value for topic modelling. Hence such stop words were also removed so the results could be more appropriate.

3.3.2 TF-IDF and n-gram approach

Using “sklearn TfidfVecorizer”, TF-IDF was applied to each document (review). This process gave some weights to the topics. After combining the overall weight for each topic across all the documents and taking its mean, following results were obtained when sorted in descending order.

```
Negative
[('ice cream',), ('wait staff',), ('fried chicken',), ('butter chicken',), ('deep fried',), ('overall experience',), ('save money',), ('mac cheese',), ('dining experience',), ('pulled pork',)]
Avg
[('ice cream',), ('pork belly',), ('fried chicken',), ('green tea',), ('deep fried',), ('mac cheese',), ('portion size',), ('overall experience',), ('friday night',), ('late night',)]
Positive
[('ice cream',), ('pork belly',), ('fried chicken',), ('deep fried',), ('pulled pork',), ('green tea',), ('service quick',), ('fish tacos',), ('fried rice',), ('chicken waffles',)]
```

Figure 3: TF-IDF topics

There were two choices while performing TF-IDF and further topic modelling with respect to n-gram approach. With initial intuition on 1-gram approach, we found that topics we get are a single word which would not provide great description. For example, if we get a positive weighted topic for restaurant as “ice”, we are not sure if its ice cream or some other product related to ice. Using bigram approach resolves this issue and provides better understanding of results. Hence, for further analysis we decided to persist with bigram approach.

3.4 Topic modelling

3.4.1 LDA Topic Modelling

The next step in implementation was to apply first topic modelling approach of LDA. Using Count Vectorizer and 2- gram approach, the multicore model training was divided among 15 worker threads. LDA iteratively assigns topics to words and at one time it would converge. In our approach, it was found that LDA converged at 10 passes and there was lesser shuffling of topics after 10^{th} pass. Even with 15 worker threads, LDA was relatively slow. The following figure shows the output probabilities of positive reviews top rated topics by LDA.

```
Positive :
[(0,
  '0.014*"fried chicken" + 0.009*"hot sauce" + 0.009*"fish chips" + '
  '0.007*"baja fish" + 0.006*"chicken katsu"'),
 (1,
  '0.013*"pork belly" + 0.013*"pulled pork" + 0.010*"green curry" + '
  '0.008*"spring rolls" + 0.008*"sweet potato"'),
 (2,
  '0.013*"pork belly" + 0.010*"mac cheese" + 0.008*"fried chicken" + '
  '0.006*"bone marrow" + 0.006*"smoked meat"'),
 (3,
  '0.031*"ice cream" + 0.014*"green tea" + 0.011*"deep fried" + 0.008*"duck '
  'fat" + 0.008*"late night"'),
```

Figure 4: LDA topics

It can be seen from the above results, that the probabilities thrown out by LDA is not interpretable for business purposes. Moreover, LDA took a lot of time to model the topics. Hence an alternate approach with more accurate results and lesser time would be more efficient for topic modelling of restaurant reviews.

3.4.2 NMF Topic Modelling

NMF naturally produces sparse representations of matrices. This makes sense in the case of documents that generally do not contain many topics. For this, each matrix "A" was broken down into two matrices "W" and "H" such that A = WH. Following results were obtained using sklearn.

```
Positive: -------------
Topic 0:   3.186*pork belly, 1.682*kimchi fries, 0.736*spice pork, 0.517*banh boys, 0.476*belly banh
Topic 1:   2.283*ice cream, 0.897*green tea, 0.510*soft serve, 0.462*black sesame, 0.295*tea ice
Topic 2:   2.231*fried chicken, 0.555*chicken waffles, 0.207*deep fried, 0.179*chicken bao, 0.141*chicken dinner
Topic 3:   0.875*green curry, 0.471*duck fat, 0.462*fish tacos, 0.457*french toast, 0.431*fish chips
Topic 4:   2.213*pulled pork, 0.621*pork sandwich, 0.308*beef brisket, 0.274*mac cheese, 0.271*bbq sauce
Average: -------------
Topic 0:   2.759*ice cream, 0.639*soft serve, 0.249*serve ice, 0.226*tea ice, 0.174*egg waffles
Topic 1:   2.575*pork belly, 0.650*kimchi fries, 0.220*ramen places, 0.218*spice pork, 0.217*pulled pork
Topic 2:   0.630*mac cheese, 0.604*fish tacos, 0.525*late night, 0.373*french toast, 0.321*deep fried
Topic 3:   2.169*fried chicken, 0.285*chicken waffles, 0.235*deep fried, 0.141*chicken dinner, 0.117*pulled pork
Topic 4:   2.118*green tea, 0.339*red bean, 0.332*tea ice, 0.242*uncle tetsu, 0.190*tea cheesecake
Negative: -------------
Topic 0:   2.584*ice cream, 0.367*soft serve, 0.363*green tea, 0.188*egg waffle, 0.176*tea ice
Topic 1:   0.830*wait staff, 0.559*half hour, 0.419*waste money, 0.377*waited hour, 0.374*took forever
Topic 2:   2.030*deep fried, 0.139*fish chips, 0.107*burger priest, 0.095*spring rolls, 0.090*fried squid
Topic 3:   2.032*butter chicken, 0.328*chicken sweet, 0.178*palak paneer, 0.167*chicken tikka, 0.132*chicken tasted
Topic 4:   1.730*fried chicken, 0.332*smoked meat, 0.303*mac cheese, 0.240*pork belly, 0.224*pulled pork
```

Figure 5:NMF topics

This approach was relatively quicker as compared to LDA and produced topic weights which can be further utilized for visualization.

3.5 Normalizing and Mapping

In order to visualize the above results, the topic weights were normalized to 1. While the relative weights remained the same, a more comparable format was obtained. The top 5 topics add up to 1 and hence we can know which topic is more highly weighed as compared to others. Average topic distributions for all reviews of a restaurant were calculated to map that restaurant to the topic space. All the dataframes were stored as csv's to further visualize the result. Hence, the implementation produced three csv's namely, topic modelled positive ratings, average ratings and negative ratings. A business intelligence tool like Tableau or Cognos BI can be further used for visualizing the results. These visualizations would help in determining the positive and negative aspects of each restaurant of Ontario and provides a scope of improvement for the restaurant owners.

4. Evaluation and Visualization

4.1 Evaluation

In this project, we performed topic modelling on more than 70,000 reviews. Two different techniques were used for topic modelling namely, LDA and NMF. The table below shows the wall time for training each model.

Table 1: Method Evaluation

Method	Wall Time
Latent Dirichlet Allocation (LDA)	3 mins 8s
Non-Negative Matrix Factorization (NMF)	349ms

As LDA requires more hardware resources (this was performed with 15 worker threads), the difference in amount of time it took to model is quite large. Moreover, the probability distribution for NMF was much more interpretable for business needs as compared to LDA.

4.2 Visualization

4.2.1 Positive reviews visualization

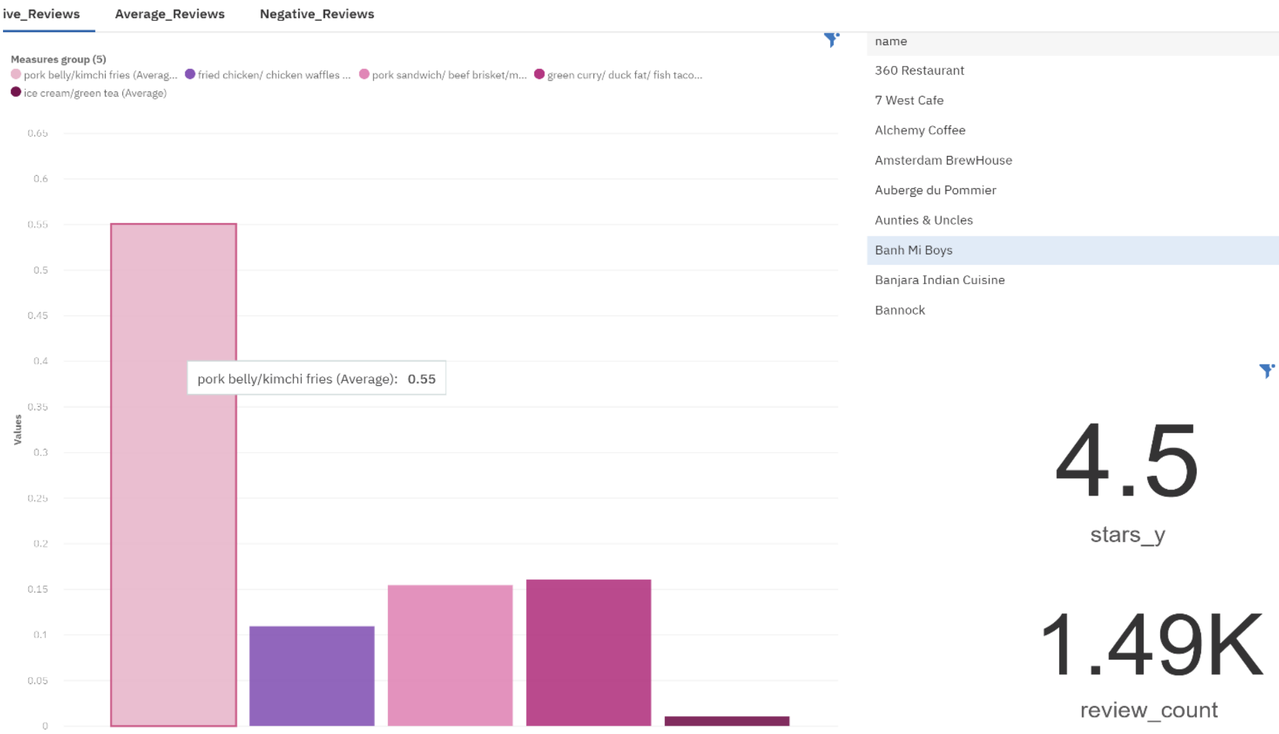


Figure 6: Positive Reviews Visualization

The above figure shows visualization for topics of positive reviews. The restaurant Banh Mi Boys has highly rated topic of pork belly in positive reviews.

4.2.2 Average reviews visualization

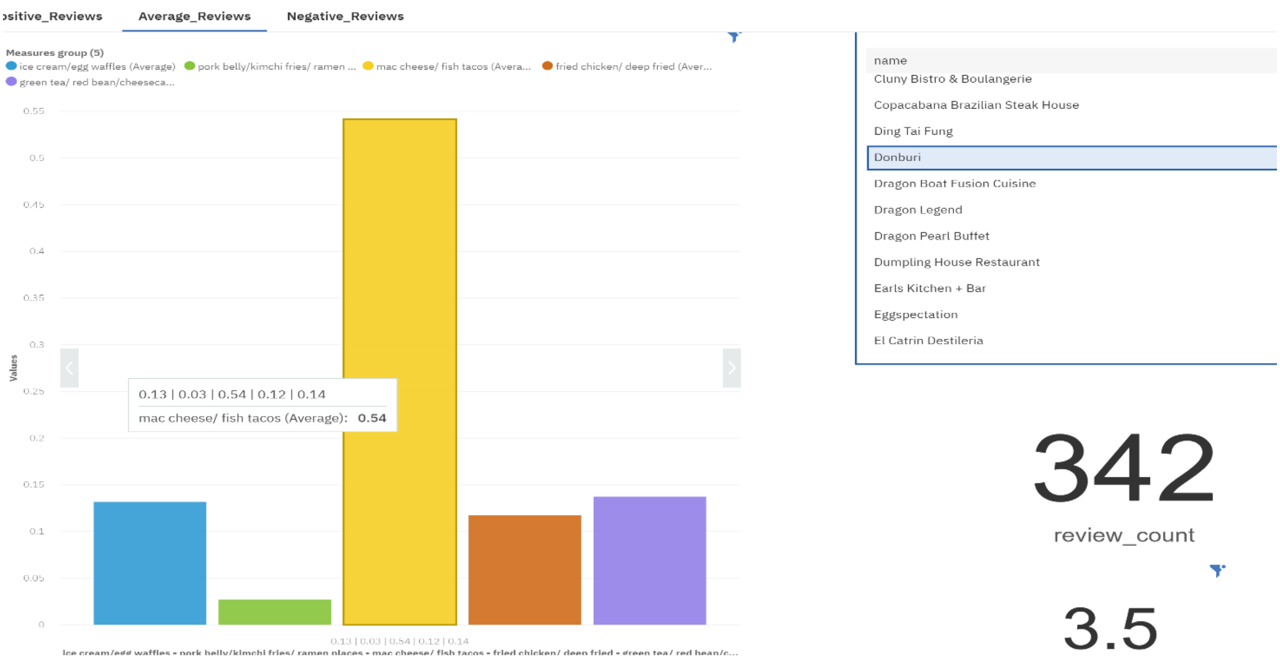
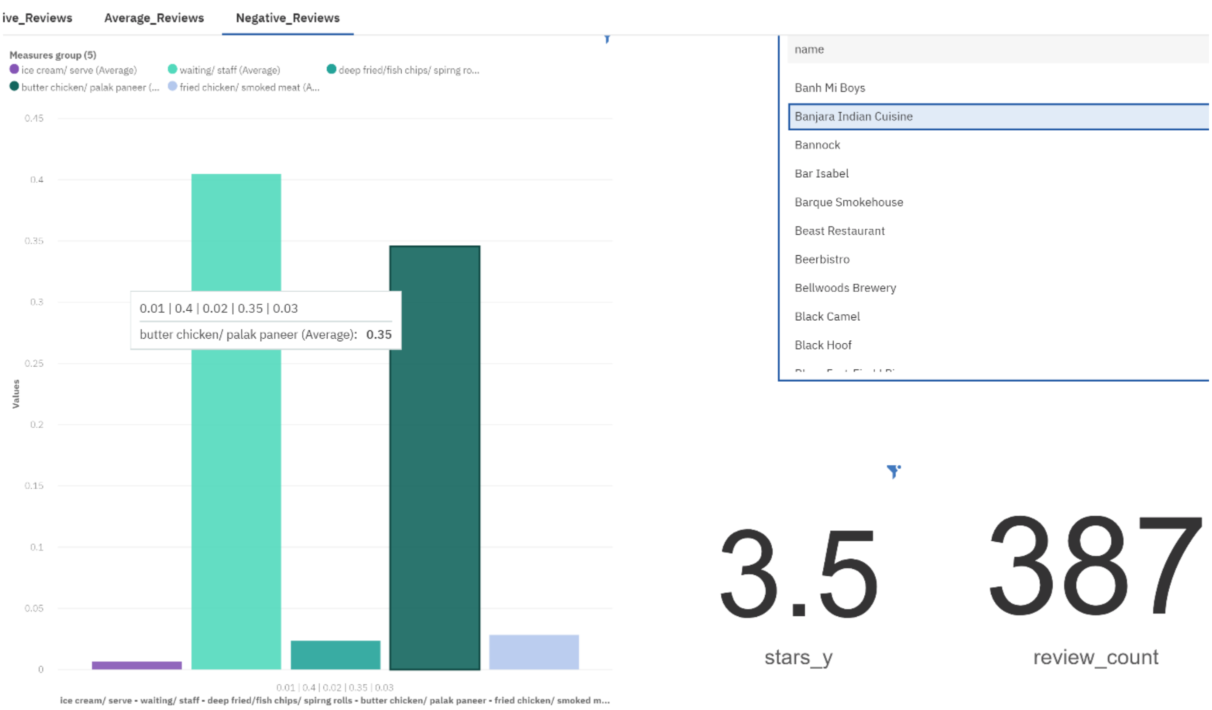


Figure 7: Average Reviews Visualization

As shown in figure, Donburi restaurant in Ontario has fish tacos as highly weighed topic among 3-star reviews of the restaurant.

4.2.3 Negative reviews visualization



As Banjara Indian cuisine has wait/staff and butter chicken as two highly weighed negative topics in reviews, it can improve on this menu item to increase their business.

5. Future Work

Topic modelling using NMF provides scope for further improvements for more accurate results. First, if we have a greater number of reviews for a restaurant than instead of taking all restaurants at once, we can model one restaurant at a time. This would help us to identify topics for individual restaurants rather than generalized topics. Secondly, instead of distributing topics directly based on rating, we can perform sentiment analysis to know the polarity of review and then divide the reviews based on intensity of the review. This both improvements can result in higher accuracy and more understandable topics.

6. Conclusion

Using topic modelling techniques of NLP, we have successfully analysed the Yelp reviews of the restaurants. Overall, it was observed that NMF performed better in terms of speed and accuracy. The visualization results showed that topic weights obtained from NMF can be practically applied to gain business insights of restaurant business.

References

[1]"Cognos Business Intelligence 10.2.2 documentation", *Ibm.com*, 2020. [Online]. Available: https://www.ibm.com/support/pages/cognos-business-intelligence-1022-documentation. [Accessed: 03- Apr- 2020].

[2] H. Parsa, J. Self, D. Njite and T. King, "Why Restaurants Fail", *Cornell Hotel and Restaurant Administration Quarterly*, vol. 46, no. 3, pp. 304-322, 2005. Available: 10.1177/0010880405275598.

[3] J. Bellini, "The No. 1 thing to consider before opening a restaurant", *CNBC*, 2020. [Online]. Available: https://www.cnbc.com/2016/01/20/heres-the-real-reason-why-most-restaurants-fail.html. [Accessed: 02- Apr- 2020].

[4] R. Fuggetta, "How to Turn Customers Into Advocates," *Restaurant Hospitality,* vol. 96, *(12),* pp. 22, 2012.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G.W. Furnas, and R. A. Harshman. “Indexing by latent semantic analysis.” Journal of the American Society of Information Science, 41(6):391407, 1990.

[6] J. Huang, S. Rogers and E. Joo, "Improving Restaurants by Extracting Subtopics from Yelp Reviews", *Ideals.illinois.edu*, 2020. [Online]. Available: https://www.ideals.illinois.edu/handle/2142/48832. [Accessed: 02- Mar- 2020].

[7] Wei Xu, Xin Liu, Yihong Gong, “Document Clustering Based on Non-negative matrix Factorization”, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR ’03), pp. 267-273, 2003.

[8] Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao, “Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews”, Center for urban science & progress, New York University, US.

[9] *Yelp.ca*, 2020. [Online]. Available: https://www.yelp.ca/halifax. [Accessed: 02- Apr-2020].