**BA 576-400 Data and Text Mining**

**Final Group Project**

**Predicting Useful Reviews on Yelp Dataset**

**Group 2**:

Eunjeong Heo

Hyomin Shin

Yi-Jen Chiang

YuKai Chen

**June 6th, 2021**

**Table of Content**

|  |  |
| --- | --- |
| 1. [Executive Summary](#_heading=h.djka19915rea) ------------------------------------------------------------------------------------ | 4 |
| 2. [Introduction](#_heading=h.p3qi2rrajf4s) ---------------------------------------------------------------------------------------------- | 5 |
| 3. [Data Exploration](#_heading=h.d3fp2y79odqo) ---------------------------------------------------------------------------------------- | 6 |
| 3.1) [Exploration of Original Datasets](#_heading=h.9vyecbez8z1j) ---------------------------------------------------------- | 6 |
| 3.2) [Merging and Sampling Datasets](#_heading=h.i3jml19e4oqs) ---------------------------------------------------------- | 6 |
| 3.3) [Exploration of Summary Statistics](#_heading=h.zc7lcihtkkwg) -------------------------------------------------------- | 7 |
| * [Image 1. Overall Summary Statistics](#bookmark=id.v3a57mzbxg52)  ------------------------------------------------ | 7 |
| * [Image 2. Summary Statistics (Median)](#bookmark=id.i0glhyyo8zeo) ---------------------------------------------- | 7 |
| 3.4) [Run Regression Model](#_heading=h.58hmg296dqaw) --------------------------------------------------------------------- | 7 |
| * [Image 3. The Regression Model Result](#bookmark=id.qth337c8ilji) ---------------------------------------------- | 8 |
| 3.5) [Exploration on Location](#_heading=h.7r5y4czat3or) -------------------------------------------------------------------- | 8 |
| * [Image 4. The Total Amount of Reviews where Useful Vote is ‘0’ by Each State](#bookmark=id.iiybmvl14zp8) | 8 |
| * [Image 5. Summary Statistics of Sample Dataset Excluding the Value ‘0’ for Useful](#bookmark=id.jmpao2ul37ee) ------------------------------------------------------------------------------------- | 9 |
| * [Image 6. Median of Sample Dataset Excluding the Value ‘0’ for Useful](#bookmark=id.n41lvvjpnwbb) ------- | 9 |
| * [Image 7. Useful Review Count Sorted by State in Three Criteria](#bookmark=id.14t43uf82cvd) --------------- | 10 |
| 4. [Data Preprocessing](#_heading=h.4v96y5gffebo): ------------------------------------------------------------------------------------ | 11 |
| * [Image 8. Screenshots of All of the Attributes from Merged Dataset (Left), and Sampling Code (Right)](#bookmark=id.yv2ndlc7yew7) ---------------------------------------------------------------- | 11 |
| 4.1) [Data Merging, Cleansing and Sampling](#_heading=h.9ptzvlvrzv1f)-------------------------------------------------- | 11 |
| 4.2) [Set a Criterion for a ‘Useful Review’](#_heading=h.wx0vqsv2es8l) ---------------------------------------------------- | 11 |
| 4.3) [RapidMiner Preprocessing -Text Preprocessing and Under Sampling](#_heading=h.9ye68bcuyecc) -------------- | 12 |
| * [Image 9. Screenshot of Main Steps](#bookmark=id.mckchpdmslly) -------------------------------------------------- | 12 |
| * [Image 10. Screenshot of Embedded Operator of Process Documents from Data](#bookmark=id.3q1laldami07) | 12 |
| * [Image 11. Screenshot of Result of the Term-Document Matrix (TF-IDF)](#bookmark=id.pn1sn8oo4cfn) | 13 |
| 5. [Text Mining & Evaluation](#_heading=h.gfe0k7s2cfut) ----------------------------------------------------------------------------- | 13 |
| * [Image 12. Screenshot of Text Mining Process](#bookmark=id.8j9v84n2v9uh) ------------------------------------- | 14 |
| * [Image 13. Screenshot of Cross Validation Embedded Process](#bookmark=id.ld0rykhjc2ro) ------------------ | 14 |
| * [Image 14. Screenshot of Final text preprocessing operators](#bookmark=id.4px8sg53a9l6) --------------------- | 15 |
| * [Image 15. Screenshot of Final text preprocessing parameters](#bookmark=id.78cg9owzwqjs) ------------------- | 15 |
| * [Table 1. Simulation Table of Text Mining Modeling](#bookmark=id.jicmquph1kt8) ------------------------------ | 17 |
| * [Table 2. Table of Featuring ‘Useful’ and ‘Not Useful’ Reviews](#bookmark=id.jd2vwwqbba21) ---------------- | 18 |
| * [Image 16. Screenshot of Final Result](#bookmark=id.5d3hr1j3uvk7) ----------------------------------------------- | 18 |
| * [Image 17. Screenshot of Final Result - ExampleSet](#bookmark=id.658m7zh902ma) ------------------------------ | 18 |
| 6. [Business Implication](#_heading=h.eztwj8846dah) ---------------------------------------------------------------------------------- | 19 |
| 7. [Further Extension -](#_heading=h.2sidy9ixbz5u)-------------------------------------------------------------------------------------- | 20 |
| 7.1) [Learning Based on the Articles](#_heading=h.v57zzmdvabx7) ---------------------------------------------------------- | 20 |
| 7.2) [Further Improvement on the Project Based on the Knowledge](#_heading=h.tq5aenlh5y3q) --------------------- | 21 |
| * [Image 18. Word Clouds Generated for Each Extracted Aspect](#bookmark=id.gy4xiul6w88p) ----------------- | 21 |
| * [Image 19. Results of Sentiment Analysis of Four Topics](#bookmark=id.dbk5o4xg9wuv) ------------------------ | 22 |
| 8. [Conclusion](#_heading=h.n23hab9lu6pd) ---------------------------------------------------------------------------------------------- | 23 |
| 9. [Reference](#_heading=h.v3mzw6biy705) ----------------------------------------------------------------------------------------------- | 24 |

### **1. Executive Summary**

Yelp online reviews have much influence on users and have an important impact on consumers’ decision making. Users tend to search for reviews from Yelp to get useful information, however, sometimes fail to get relevant, reliable and trustworthy reviews. The reasons are high rating reviews tend to be outdated and latest reviews tend to have low or none of useful votes, and some businesses do not have any reviews that are voted as useful. This is problematic because users and businesses cannot define which is the most relevant review when searching on the review, have difficulty in which reviews they should value the most, and if so, users need to look through lots of reviews to find the newest and most relevant reviews. Hereby, this project aims to predict which review will be rated as useful by using classification algorithm Naïve-Bayes method in order to solve the problems generated by lack of useful reviews. The model was evaluated by accuracy which is the most important measurement for classification models. The project went through trials of adjusting different techniques and parameters, and got 94.5% of accuracy for predicting the useful reviews. As a result, the project could further identify common patterns of useful reviews and the differences between useful and non-useful. The useful reviews tend to include more sentences, detailed experience with examples, use of comparative adjectives and exclamation marks, and less grammatical and spelling errors. Whereas, non-useful reviews had shorter sentences, toneless explanations with more grammatical and spelling errors. With these interesting findings, users and businesses can easily get access to useful reviews, and identifying the useful reviews will enhance effectiveness and efficiency of decision making and business operations. In addition, the findings can be used in the Yelp algorithm to significantly differentiate its user-experience from other review platforms.

**2. Introduction**

Yelp is one of the most trusted, influential and useful platforms for making a purchase decision (Belt, 2017). According to Wiideman Consulting Group, 60% of respondents answered that the primary reason for using or visiting Yelp is to read reviews (Wiideman, 2015). Like 91% of consumers regarding online reviews as personal recommendations (Marie, 2021), many users view reviews with a ‘useful’ recommendation tool provided by Yelp to avoid untrustworthy or fake reviews and use it for decision making.

However, the problem is that ‘useful’ votes reviews do not exist for all businesses. Some businesses do not have reviews with useful votes making users unaware of which reviews are credible. In addition, reviews on Yelp typically cannot achieve useful votes at a fast pace. It needs several weeks or months in order to achieve lots of useful votes. Therefore, results in fewer latest useful reviews and by the time when the latest review becomes a useful review it might be outdated and not be informative for the users. Lack of reviews with useful votes can be problematic for users and business owners. Users cannot get exact, relevant and reliable recommendations for decision-making. 29% of consumers make purchases on the same day, and 89% make purchases on the same week they view reviews but without useful reviews users cannot obtain reliable information for purchase decisions (Dato). For businesses, without useful reviews it may lose the opportunity to make customers spend 31% more on businesses (Dato), and furthermore lose the opportunity to get higher profit. Moreover, without knowing the useful reviews, businesses will have difficulties predicting why customers are coming to or what customers are expecting from the business. Since the reviews are not ranked as useful, businesses cannot get insights on its product or service to prepare for operation, promotion and improvements.

Therefore, predicting useful reviews by using useful votes is necessary to provide insights to the users and businesses in order to get benefit from reviews. By classifying the useful reviews, our project aims to define useful reviews for users to get the most helpful review without being deluded. For businesses, our project seeks to provide the results for businesses to prepare for business activities in advance based on the prediction of useful reviews.

**3. Data Exploration**

To define and predict the useful reviews, the datasets needed in this project are 'yelp\_review, 'yelp\_business', and 'yelp\_user' datasets. In order to find out the correlation between the useful review and location or useful reviews and any variables. For instance, funny, cool reviews, business review count, business stars and review stars.

**3.1) Exploration of Original Datasets**

To explore the data of the three original datasets, yelp\_review, yelp\_business, and yelp\_user.

* Original “yelp\_review” dataset statistics

Based on the core problem we expect to solve, the “yelp\_review” is one of the main datasets that is needed for identifying useful reviews. The review dataset is the most important since it is the fundamental source for detecting the useful reviews. The “Yelp\_review” dataset contains 5,261,608 rows, 9 columns, and has records from July 22, 2004 to December 11, 2017. In addition, some of the text is not English and some of the values are missing.

* Original “yelp\_business” dataset statistics

The “yelp\_business” dataset contains business location information which may be a significant attribute when determining useful reviews. Moreover, the business table is needed in order to find out relevant information associated with the review business\_id. The “yelp\_business” dataset contains 174,134 rows, and 13 attributes.

* Original “yelp\_user” dataset statistics

The “yelp\_user” dataset is needed to match the user\_id to the review dataset user\_id and see which of the users have written down useful reviews and see if there is any similarity among those users. The screenshot below is a portion of the summary statistics, in the table it explains how many friends the user has or which types and how many compliments a user got from other users. The compliments attribute’ averages are in the range of 0 to 3 and have a big difference in the range of min and max.

**3.2) Merging and Sampling Datasets**

After exploring the original dataset, as recommended, our team went through Python to randomly sample around 13,000 reviews of the original “yelp\_review” dataset for the project. After, further combined the three datasets into one file in order to see the relationships between useful and other variables.

**3.3) Exploration of Summary Statistics**

To find out about the sample dataset, we summarized the attributes by summary statistics. By exploring the summary, our team identified the values of mean, std, min, max and median in the data and whether the data is skewed or not. The following image is the overall summary statistics of important attributes in the dataset.

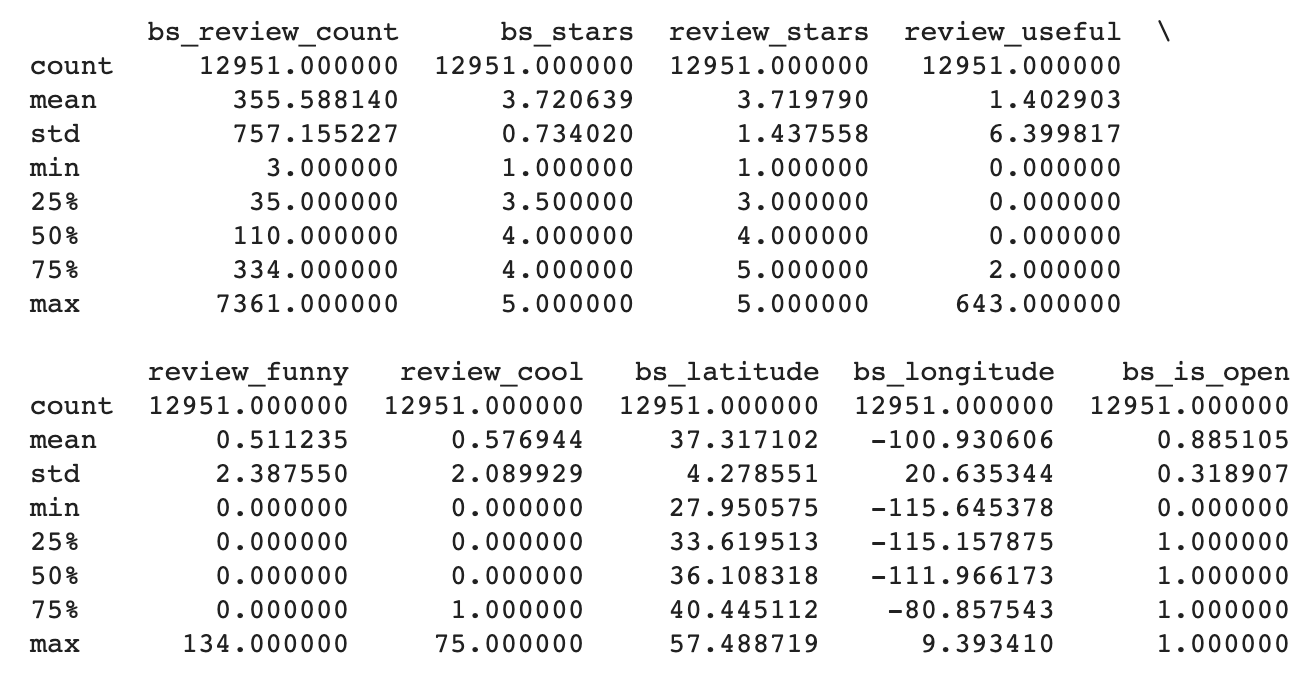


Image 1. Overall Summary Statistics

In addition, looked at the median for each attribute. By looking at the median, the review\_useful had a median value of 0. This indicated that data mainly had the value of 0 and data may be skewed to the zero.

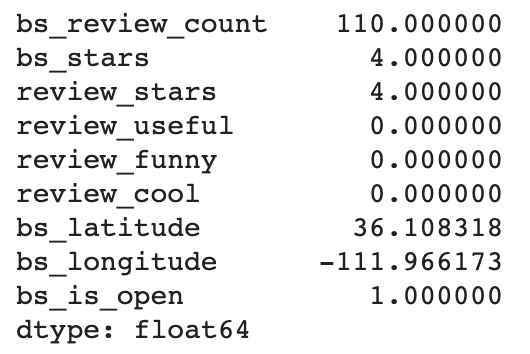


Image 2. Summary Statistics (Median)

**3.4) Run Regression Model**

Next, in order to figure out the correlation between useful and other variables our team ran a regression model. We identified the correlation between useful and ‘bs\_review\_count’, ‘bs\_stars’, ‘review\_stars’, ‘funny’, and ‘cool’. The business review count, and stars had negative correlation. Despite the fact that the regression model shows funny and cool variables’ relationships being positive, it is not strongly correlated. Therefore, our group concluded that for the project, we should focus on the review text and useful attributes in the review dataset.

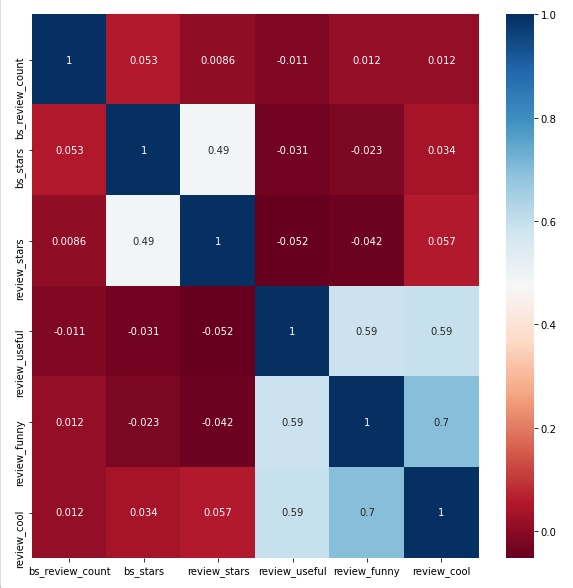


Image 3. The Regression Model Result

**3.5) Exploration on Location**

Our project further explored the relationship with useful attribute with the location to decide whether to limit the location to a specific area or not. Through the statistics summary result, we found out that at least 50% of reviews have zero useful votes. To examine the outcome, we further analyzed to see the number of reviews that got zero useful votes in each state and compared with the total number of reviews of each state.

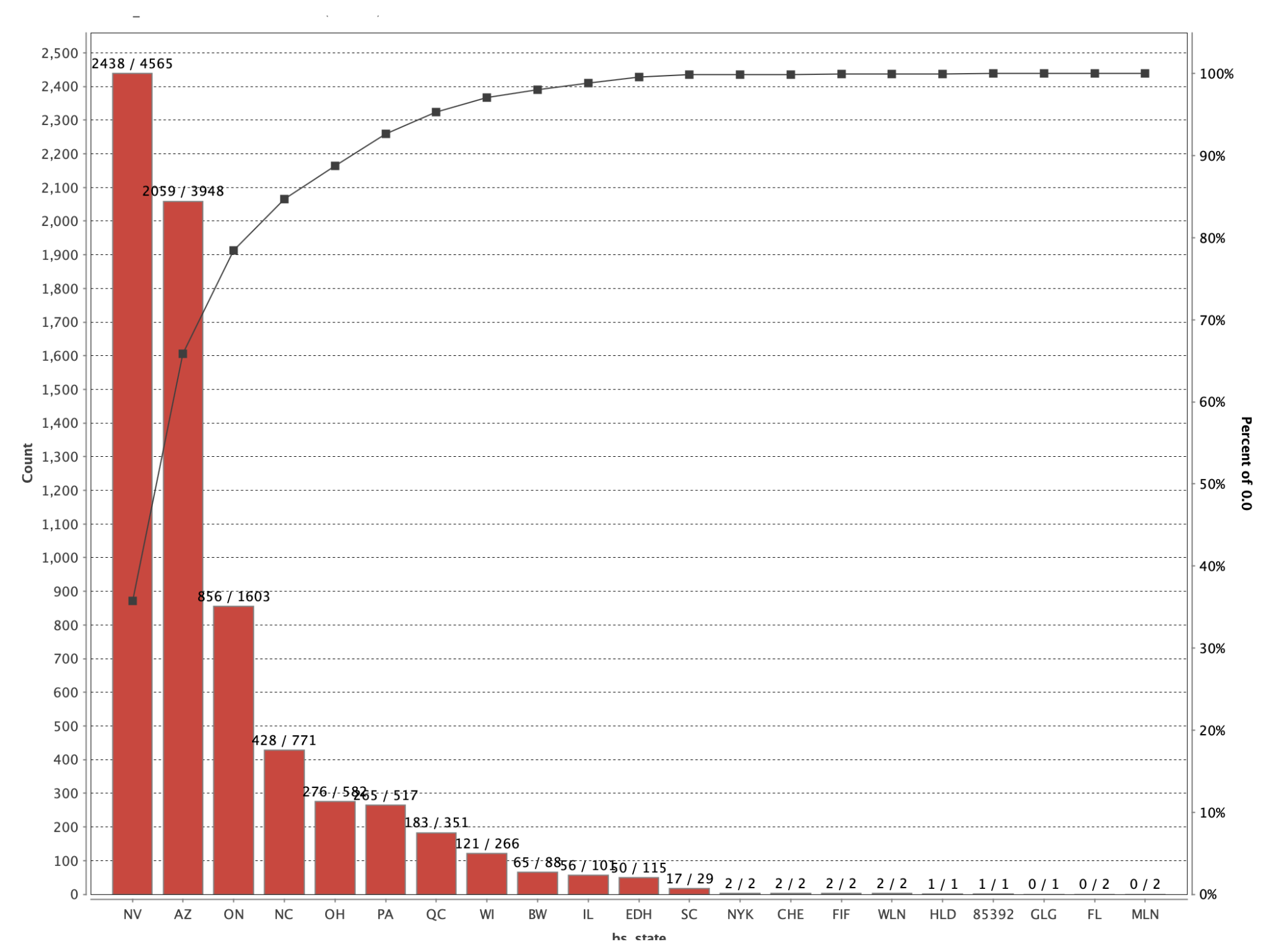


Image 4. The Total Amount of Reviews where Useful Vote is ‘0’ by Each State

Looking at the ‘Image 4’, it indicates that all states have evenly distributed with about the same ratio of 5:5 to reviews that have zero useful votes and have over zero useful votes. Therefore, there is no need to narrow down to a specific area.

In order to examine it deeply, we decided to exclude the reviews with zero votes of useful attributes and compare the summary statistics outcome with previous statistics that included the value of zero.

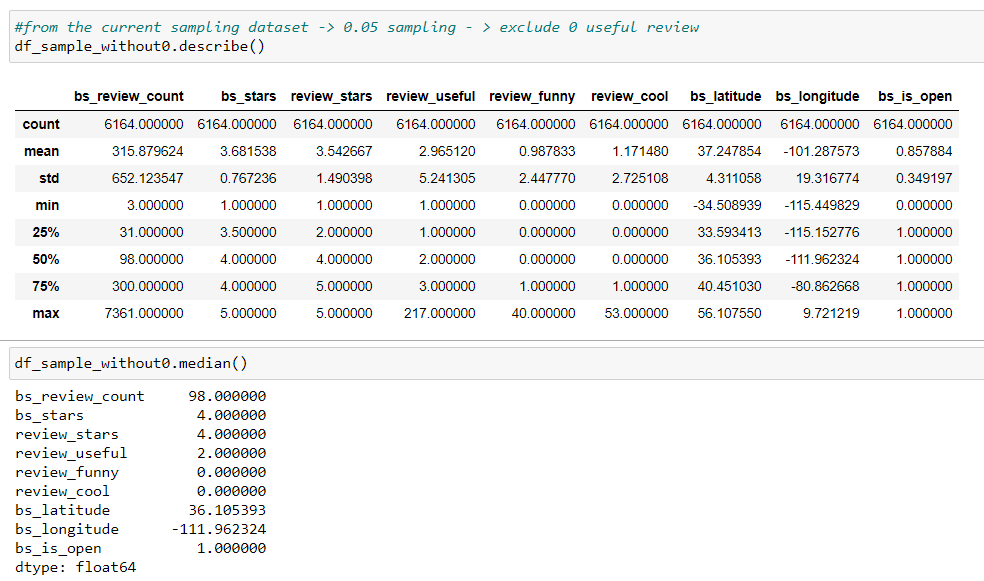
****

Image 5. Summary Statistics of Sample Dataset Excluding the Value ‘0’ for Useful

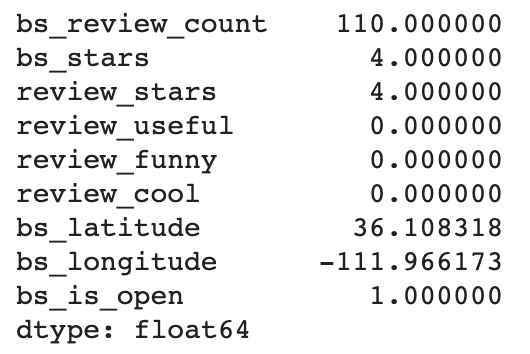


Image 6. Median of Sample Dataset Excluding the Value ‘0’ for Useful

After looking at the statistics, our team did a visualization of the useful review count sorted by state.

For visualization, we looked with three criteria of median, mean and standard deviation.

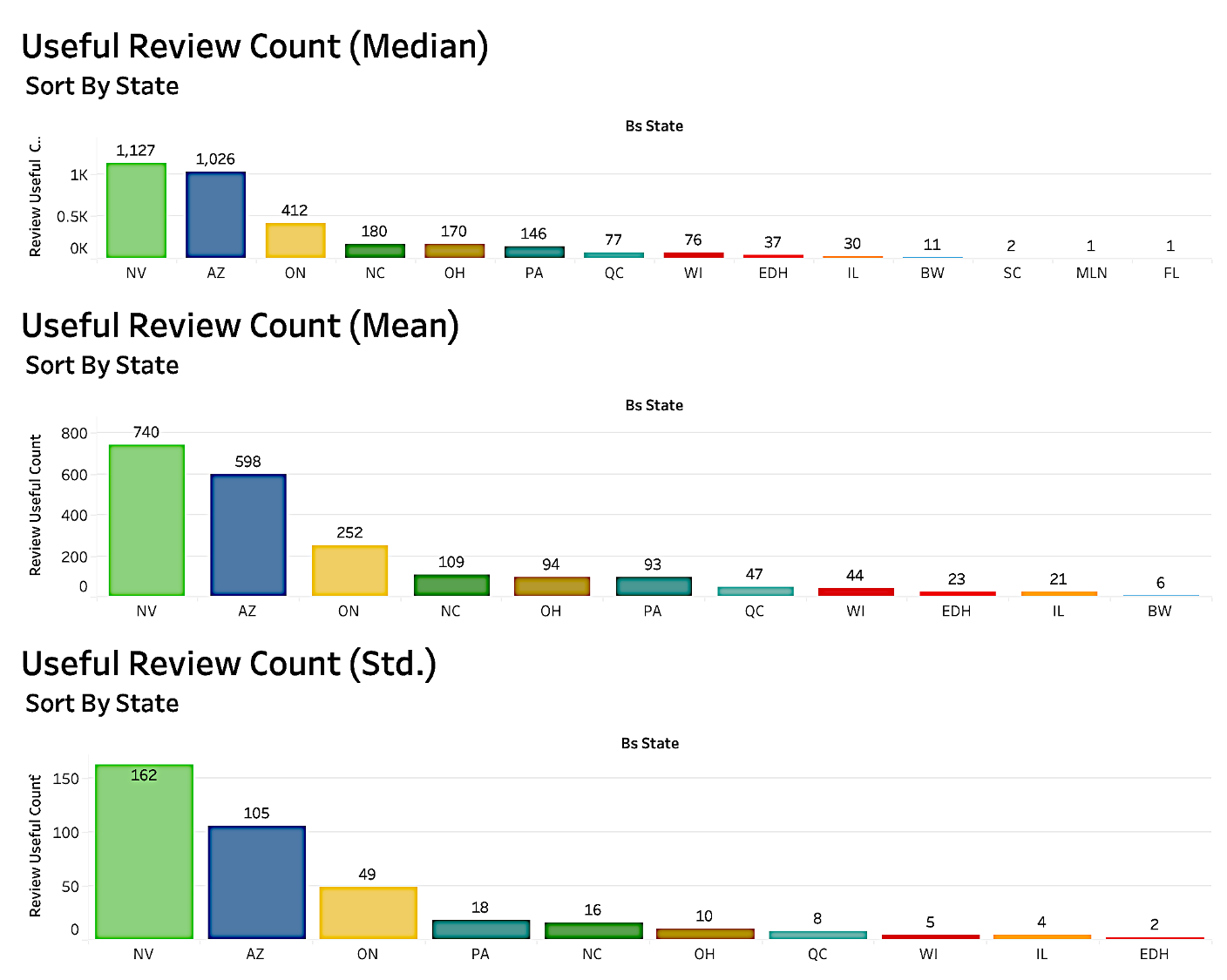


Image 7. Useful Review Count Sorted by State in Three Criteria

* Criteria ‘median’:

I. Contains 3,296 reviews equal or more than median.

II. Including 14 states, and the top five states on the list are NV, AZ, ON, NC, OH.

* Criteria ‘mean’:

I. Contains 2,024 reviews equal or more than mean.

II. Including 11 states, and the top three states on the list are NV, AZ, ON, NC, OH.

* Criteria ‘Standard deviation’:

I. Contains 379 reviews equal or more than Standard deviation.

II. Contains 10 states, and the top three states on the list are NV, AZ, ON, PA, NC.

**4. Data Preprocessing**

Our team initially decided to use 3 datasets, ‘user’, ‘business’, and ‘review’, to see if there is more information we could obtain besides our main dataset ‘review’ to validate useful reviews. For example, we expected to see some useful reviews’ patterns amongst states or correlation between business rating and useful reviews, etc., therefore, ended up merging the three datasets and randomly sampled the data to be able to make it run in RapidMiner. We investigated whether the dataset has duplicates and null values using Python, after dropping them, and moved on to the sampling phase.

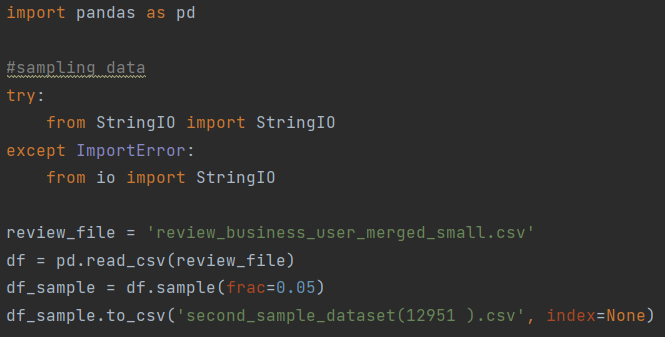


Image 8. Screenshots of All of the Attributes from Merged Dataset (Left), and Sampling Code (Right)

**4.1) Data Merging, Cleansing and Sampling**

As predicting the potential useful reviews is the initial goal of our project, the business and user data sets were merged into left-join based on the review sample data set. We then identified some duplicates and missing values and removed them by using python-pandas, based on the observation on the data exploration phase.

The 'sampling' process, which organizes parts of existing data (population) and makes it as optimal input data, is the most fundamental step in the data preprocessing process. Sampling refers to the process of extracting a certain number from a population. It is a statistical procedure related to the selection of individual observations, which helps to make statistical inferences about the population. As our group’s goal is to identify potential useful reviews, we concluded around 13,000 independent observations would be enough.

**4.2) Set a Criterion for a ‘Useful Review’**

After exploring the data ‘second\_sample\_dataset(12951).csv’, to create a training dataset, we concluded there are 3 different potential criteria to identify useful reviews; mean, median and standard deviation. As there are too many 0 ‘useful votes’ in our dataset, we excluded them first to prevent diluting the criteria of useful reviews and generated 3 different binary criteria datasets (1: useful, 0: not useful) based on each dataset’s defined criteria:

‘second\_sample\_dataset(12951)\_mean.csv’; 1: useful mean > =3, 0: useful mean < 3

‘second\_sample\_dataset(12951)\_median.csv’; 1: useful median >= 2,0: useful median < 2

‘second\_sample\_dataset(12951)\_std.csv’; 1: useful std >=5, 0:useful std < 5

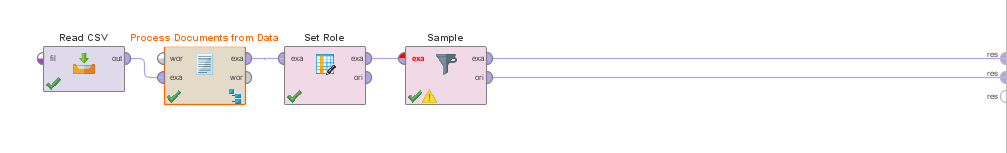
**4.3) RapidMiner Preprocessing – Text Preprocessing and Under Sampling**

Image 9. Screenshot of Main Steps

In RapidMiner, we used the ‘Read CSV’ operator to import the merged dataset and classified 1 and 0 in the sampling dataset.

Initially, before applying the text mining algorithm, our project used the pruning method. The prune method (prune below absolute: 5, prune above absolute: 500) helped the dataset to improve acquiring the more meaningful text and also initially applied TF-IDF. It is a method to calculate the weight of words in a document by finding words that do not appear in other documents but only in certain documents, TF, and finding a word frequency that shows how often a particular word appears in a particular document, DF. Therefore, the result shows a value that indicates how often a particular word appears in a whole document.

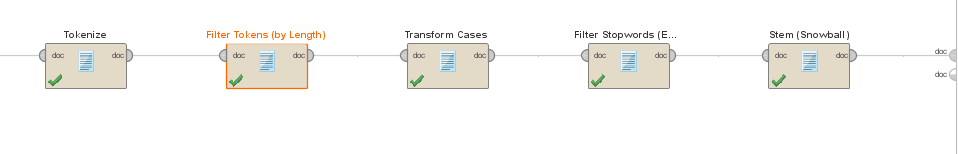


Image 10. Screenshot of Embedded Operator of Process Documents from Data

Under the embedded process of ‘Process Documents from Data’, we initially ran ‘Tokenization’, ‘Filter Stopwords (by English)’, ‘Transform Cases’, ‘Filter Tokens' (by length’), and ‘Stem (Snowball)’ operators. ‘Filter Tokens (by POS)’ was excluded as the process took more than 10 minutes to complete. ‘Tokenization’ is the most basic step of text preprocessing as the combination of texts should be broken down in the individual work. The text can be divided into words, spelling, phrases, sentences, or paragraphs. ‘Filter Tokens (by length)’ removes tokens shorter or longer than specified character lengths. It is a word length trimming process to show the only information needed and we set the parameters to a minimum of 3 and a maximum of 25. Next, we transformed all cases to ‘lower-case’ and normalized the various types of the same meaning words to one unified ‘stem’ word. Using ‘Stop words’ then helps the text to remove words that form a non-linguistic view or do not carry any meaningful information itself. This operator helps the process to perform better.

The given dataset has much more class of 0 (not useful) than 1 which typically results in degraded performance of the learned model due to the dominant influence of the observed category of data. Therefore, we under-sampled the given data set by matching the number of samples based on the low-rate data value by using the operator ‘Sample’ in the RapidMiner.

After completing the text-preprocessing and under-sampling, the result is as follows:

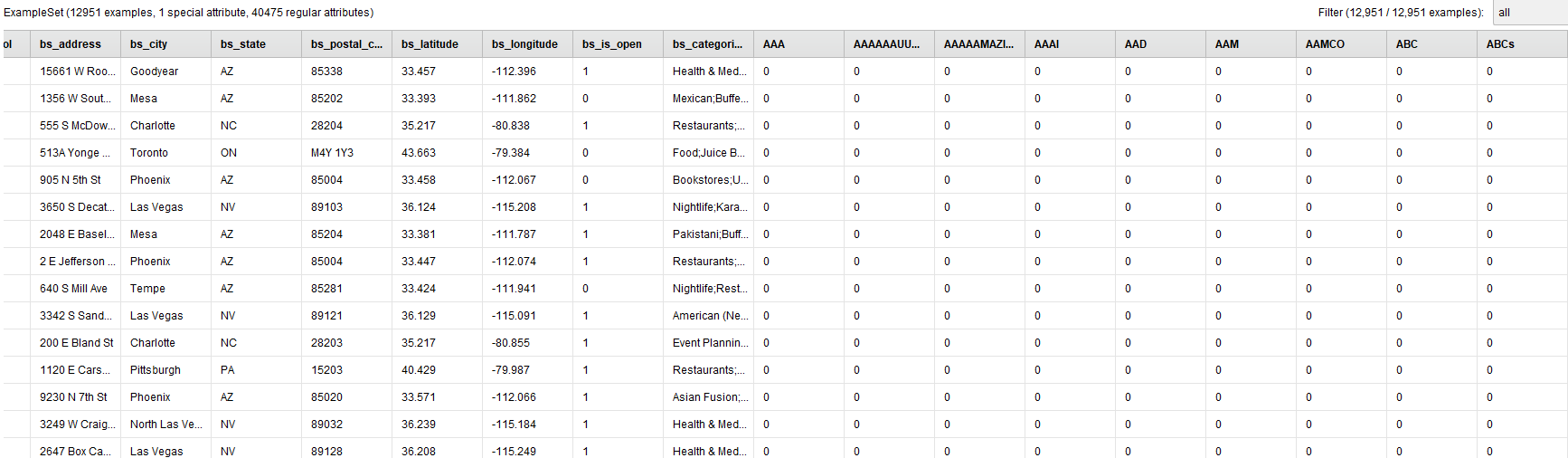


Image 11. Screenshot of Result of the Term-Document Matrix (TF-IDF)

**5. Text Mining & Evaluation**

Normally, accuracy is the most intuitive measure for evaluating classification models. It is an obvious metric, which measures all the correctly identified cases. Accuracy is most used when all the classes are equally important. However, when the data is unbalanced, precision, recall and F1-score could be better. Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives and recall represents the percentage of the positive result, how often does it predict correctly. Lastly, F1-score is the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account.

The given dataset has the imbalanced data between class 1 and 0 in 1:5 ratio. If the cost of making a misclassification was higher for one class than the others or the benefit of making a correct classification was higher for one class than the others, we could have used the imbalanced dataset itself to use precision, recall and f-1 score for evaluating our prediction model. However, in our case, we are classifying ‘useful’ and ‘not useful’ that are equally important because not only predicting ‘1(useful)’ brings extra profit by preparing the strategy, but also mis-classified ‘0(not useful)’ could cause a financial damage by ruining yelp user’s experience. Therefore, in the preprocessing stage, we under-sampled the imbalanced data because if the amount of data between 1 and 0 is too large, only the learning model of the dominant class takes place.

To identify the best training dataset amongst mean, median, and standard deviation, each member has been assigned one dataset and tried several different models including decision trees, naïve bayes and KNN, and pre-processing operators and parameters such as stem, n-gram, filter by length, prune, type of vector creation, oversampling parameters, etc. Consequently, the dataset with median criteria’s average accuracy was 80%s to early 90%s, the standard deviation criteria were 70%s, yet the mean criteria got between 80% to 95%s, therefore, we narrowed the dataset down to the ‘second\_sample\_dataset(12951)\_mean.csv’.

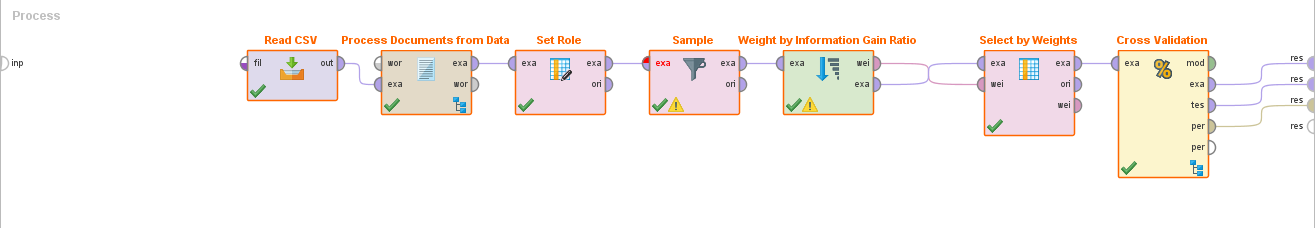


Image 12. Screenshot of Text Mining Process

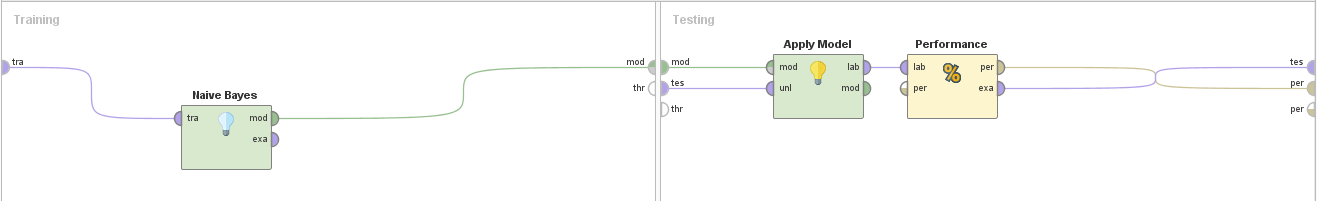


Image 13. Screenshot of Cross Validation Embedded Process

During the dataset modeling exploration phase, regardless of what data set we used, decision trees or KNN algorithms brought relatively lower accuracy. Naïve-Bayes always brought the highest accuracy. Decision tree is a supervised machine learning where the data continuously split according to a certain parameter. After modeling with several different operators, we concluded a ‘decision tree’ model is not appropriate for identifying useful reviews, as large trees can be difficult to interpret and the decisions it makes may seem counterintuitive. In addition, it is easy to overfit, the decision boundary restricted being parallel to attribute axes, and decision tree models are often biased toward splits on features having many levels (Chakure, 2019).

KNN algorithm is used to classify by finding the K nearest matches in training data and then using the label of closest matches to predict. Traditionally, distance such as Euclidean is used to find the closest match. It assigns a document to the class that contains the majority of its top k nearest neighbors (most similar documents). However, similar to decision trees, in large datasets the value of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm and does not work well with high dimensions (Chatterjee, 2020). Therefore, in our case, the KNN algorithm was also not the best fit modeling.

Finally, Naive Bayes algorithm, which we ended up using in our final model, is based on the Bayes theorem and assumes conditional independence. It is simple, fast, and very effective, especially, it works well regardless of the size of the data when training the model, therefore, estimation probabilities for prediction could be easily obtained. At this point, we decided to use the ‘Weight by Information Gain Ratio’ operator instead of ‘Weight by Information Gain’ with top k 3000 weight relation because it calculates the weight of attributes with respect to the label attribute by using the information gain ratio. The higher the weight of an attribute, the more relevant it is considered.

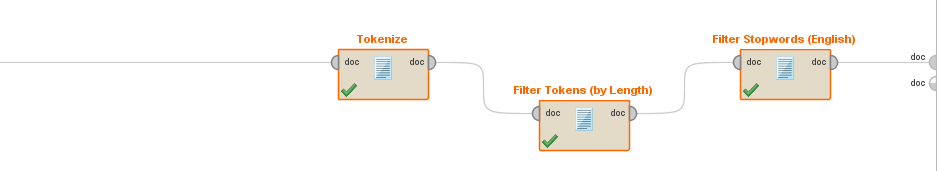


Image 14. Screenshot of Final text preprocessing operators

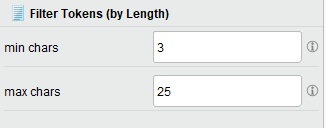
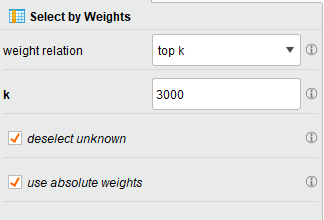
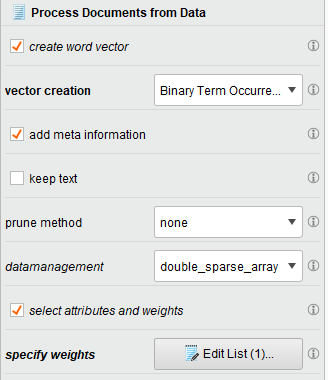


Image 15. Screenshot of Final text preprocessing parameters

After running through several different operators and parameters, the final model we ended up with are Binary Term Occurrences vector creation, Tokenize, Filter Tokens (by length) with minimum char 3 and maximum char 25, and Filter Stop words. Binary Term Occurrences change the number of word occurrences in the document to only 1 and 0 which simply indicates that there is at least one occurrence of the word token "mining".

|  |  |  |
| --- | --- | --- |
| Operators | Parameter | Accuracy |
| Tokenize, Filter Tokens (by length), Filter Stop Words, Transform Cases, Stem (Snowball) | TF-IDF,  min char: 3  max char: 25,  under sampling (1000,1000) | 83.45% (top k 1000)  84.04% (top k 2000)  83.65% (top k 3000)  79.25% (top k 4000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words, Transform Cases | Binary Term,  min char: 3  max char: 25,  Prune 4, 80 | 77.28% (top k 1000)  81.41% (top k 2000)  84.27% (top k 3000)  84.26% (top k 4000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words, Stem (Snowball) | Binary Term,  min char: 3  max char: 25,  under sampling (1000,1000) | 89.4% (top k: 3000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words, Stem (Snowball) | Binary Term,  min char: 3, 4  max char: 25, 20,  under sampling (1000,1000) | 94.25% (top k: 3000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words, Stem (Snowball) | Binary Term,  min char: 3, 4  max char: 25, 20,  Prune (3,100)  under sampling (1000,1000) | 94.35% (top k: 3000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words | Binary Term,  min char: 3, 4  max char: 25, 20,  Prune (5,200)  under sampling (1000,1000) | 92.9% (top k: 3000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words | Binary Term,  min char: 3  max char: 25  under sampling (800,800) | 92.11% (top k: 3000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words | Binary Term,  min char: 3  max char: 25  under sampling (2000,2000) | 91.58% (top k: 3000) |
| Tokenize, Filter Tokens by Length, Filter Stop Words | Binary Term,  min char: 3  max char: 25  under sampling (1000,1000) | 94.50% (top k: 3000) |

Table 1. Simulation Table of Text Mining Modeling

Compared to the first phase of text pre-processing, ‘Transform Case’ and ‘Stem (Snowball)’ operators are excluded because for prediction models to identify useful reviews, ‘past tense’, ‘comparative’, ‘superlative’ or ‘capital letter’ does play a significant role. ‘Stemming’ and ‘Transform Cases’ could be useful in a case when there is no difference between different forms of the text. For example, if the main goal of text mining is for identifying hot topics from the documents, transforming all cases to one type and normalizing the various types of the same meaning words to one unified ‘Stem’ word might be more useful to make the model run much more efficient by giving the model to save some time and more chances to try other operators out. However, after trials, we figured out that the reviews in the dataset contained capital letters and different types of adjectives that presents unique patterns for useful reviews as follows:

“This is my new go to wax spot. Jennifer was AMAZING! She was very kind and friendly and it was clear that it was genuine. She has earned a new customer. (…) She was great and I will be returning in four weeks. Thanks Jennifer! (useful: 4, funny: 0, cool: 3)”

“WORST experience ever and RUINED MY LIFE!! No bedside manner. NONE! It's a MONEY MAKING SCAM! DO YOUR RESEARCH! (…) Just BEWARE! DO YOUR RESEARCH I WISH I HAD. (useful: 12, funny: 2, cool: 0)”

Some of the classified useful reviews emphasized their feelings with capital letters and used past tense to describe their experiences from the business. Furthermore, a lot of useful reviews used comparative or superlative adjectives such as ‘slower’, ‘smoother’, ‘lighter’, ‘heavier’, ‘greatest’, etc. Therefore, if those features are eliminated by ‘stemming’ or ‘transform cases’ operators, the accuracy of the model decreases.

|  |  |  |
| --- | --- | --- |
|  | Useful | Not useful |
| Length | More than 1 paragraph | Shorter sentences |
| Contents | Detail-oriented experiences with taking specific examples | Not enough contents to understand the business |
| Expression | Frequent use of adjectives, comparatives and many exclamation marks | Monotonous explanation |
| Grammar | Less grammatical and spelling errors | More grammatical and spelling errors |

Table 2. Table of Featuring ‘Useful’ and ‘Not Useful’ Reviews

Consequently, the final and highest accuracy we got is 94. 50% as follows:

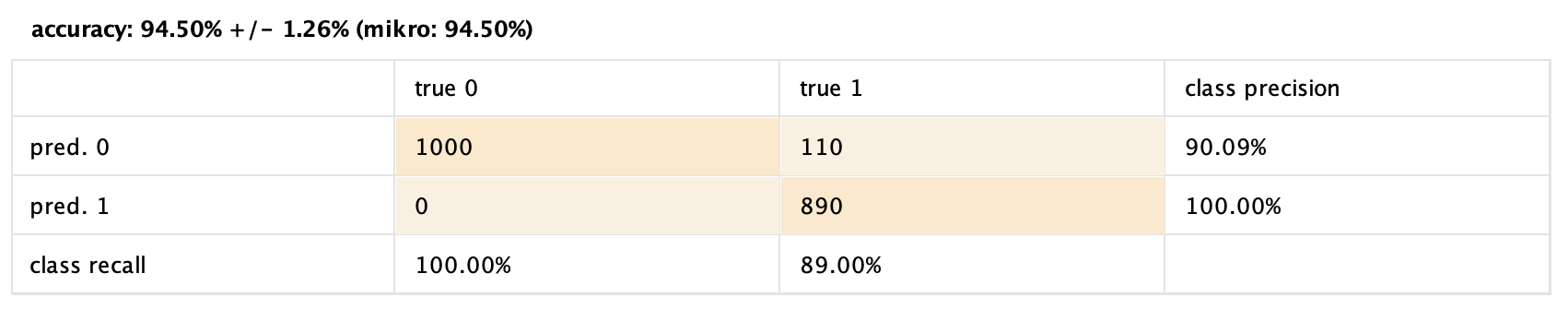


Image 16. Screenshot of Final Result

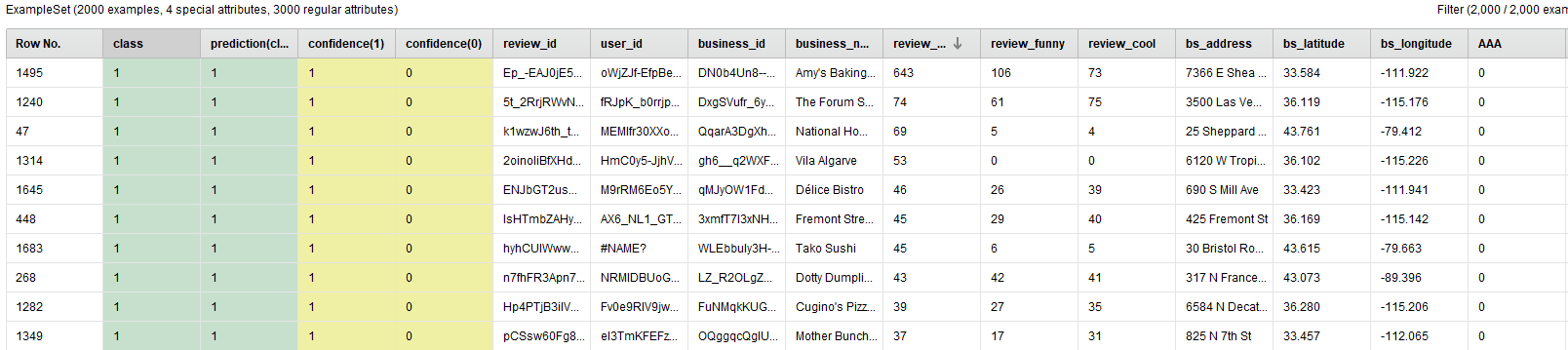


Image 17. Screenshot of Final Result – ExampleSet

**6. Business Implications**

Based on the findings from the result, users and businesses can benefit by knowing the common characteristics of ‘useful’ and ‘not useful’ reviews.

For users, understanding the characteristics of useful reviews will allow them to make effective and efficient purchasing decisions. For example, users can identify useful reviews by seeing whether the review contains common features of comparative adjectives or not. If comparative adjectives such as lighter, heavier, friendlier, superior, thicker and etc are contained, users can easily find out that the review is useful. With this common feature, users can acknowledge which comment is reliable and trustworthy. Knowing the credibility of the review, users can make a reasoned decision and get great deals. In addition, by knowing the common features used in the useful reviews, users can even identify the likelihood of the review being voted as useful or not in the future. For instance, common words of useful reviews such as absolutely, ultimately, genuinely, and effectively give users a tone of confidence about the review, therefore, it is likely to be ranked as high. When customers are confronted in a situation where there aren't any useful votes, users can apply the criteria of useful reviews and predictions by themselves. Moreover, these findings not only prevent users from believing in fake reviews or non-useful reviews but also users can utilize the common words to generate useful reviews by themselves.

For businesses, knowing the characteristics of useful reviews can facilitate more efficient business operations. It provides an opportunity to monitor and assess what the business is doing well and what they are not, and also provides guidance to prepare for efficient management. Reviews that are ranked as useful tend to have a big influence on users and therefore, the product or service mentioned in the useful reviews often becomes one of the products that yields the most profit. For example, reviews that contain comparative adjectives mention the same product as positive; the company can use that information to plan ahead. Businesses can retain supplies or inventories for a specific product mentioned in the useful reviews in advance to prevent out of stock. Likewise, with the common words, businesses can predict useful reviews in advance and utilize the information in order to prepare for operations and promotions. Businesses can improve the quality or revise their product or services based on useful reviews. In addition, businesses can plan for promotions based on the useful reviews to maximize customers' satisfaction and need. Moreover, businesses can learn from useful reviews. If a company does not have any reviews or useful reviews, the company can learn from other businesses’ useful reviews by utilizing which factors lead other businesses to success. By doing so, it can take advantage of implementing a similar atmosphere, service, menu, offerings, or promotions into their own businesses to yield better results.

Furthermore, based on the findings, Yelp could create an algorithm for recommending potential useful reviews based on our proposed useful reviews feature. Showing the potential relevant useful reviews on the top like YouTube or Amazon recommendation systems can be very beneficial to users and businesses to recognize up-to-date informative useful reviews.

**7. Further Extensions**

To further extend our project, we read through ‘Finding the reviews on yelp that actually matter to me: Innovative approach of improving recommender systems’, and ‘Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp’ articles and learned that to apply the Latent Dirichlet allocation (LDA) algorithm in our project could improve the analysis result application.

**7.1) Learnings Based on the Articles**

Nowadays, recommender mechanisms on the major review websites in the hospitality and tourism industry have excessively simple and limited options (Luo, Tang, Kim, Wang, 2020). Yelp, for example, recommends only sorting by ‘Newest First’, ‘Oldest First’, ‘Highest Rated’, ‘Lowest Rated', and ‘Elites’. These kinds of filters are not effective for some users who want to sort by a specific criterion other than those provided by date and ratings. For example, customers who want to make the right purchase decision based on the reviews by the criteria of price, menu, or product/service or etc., might have difficulty in the decision-making process.

Through the two research articles mentioned above, we learned that Latent Dirichlet allocation (LDA) is an extensively applied technique that is similar to the K-mean model but better to give more realistic results than the K-mean for topic assignment. Latent Dirichlet allocation (LDA) is an approach used in topic modeling based on probabilistic vectors of words, which indicate their relevance to the text corpus (Nodus Labs, 2018). As we learned from the text mining class, the K-mean is partitioning the N documents in K disjoint clusters. On the other hand, the LDA assigns a document to a mixture of topics. Thus, each document is specified by one or more topics. To sum up, in order to make the model work, LDA needs to pre-identify how many headings it’s searching for beforehand and could further assign the review data into several aspects to better classify words in the reviews.

**7.2) Further Improvement on the Project Based on the Knowledge**

In our project, we were only able to predict whether the new review will be useful or not. Based on the classified useful reviews, businesses needed to manually go through and refer to those reviews to predict the concerns or needs of the customers. However, after reading the research articles, we noticed that our project has room for improvements to provide a better user-experience for users and businesses by adding more specific criteria of sort options with useful votes.

From the user's perspective, not all recommended useful reviews are valuable as each user has different expectations from the businesses. Even though the dominant reviews of a restaurant are about dining atmosphere or the food quality, if the users put more value on price or products/services, the current useful reviews sorting system might not be valuable for them. Therefore, the review page may not provide the best review for the users when they are looking for helpful and useful reviews to make the right decision.

From the business’s point of view, the businesses can refer to useful reviews to realize what customers like or dislike to improve the business activities. However, it is difficult for the business to look through all of the useful reviews posted and find users' expectations exactly only with the date and ratings sorting system. On the other hand, by using the LDA method, the model could further show useful reviews by specific categories that customers value about. In the result, finding the specific expectation of each business industry and implementing new options will enhance the accuracy of the useful reviews. Consequently, we believe that the below method would optimize the application of the review board.

In order to learn how to apply the LDA method and get the accuracy result, we found an example for the restaurant category. Based on the LDA analysis of unsupervised learning algorithms, we need to define how many aspects(K) we expect to find in the reviews. In the example, it defined the K value as four which means there are four aspects that need to be considered. Based on the observation, it defined the four aspects which are taste/food, experience, value and location.



Image 18. Word Clouds Generated for Each Extracted Aspect

After running the aspect extraction, the outcome of LDA in the restaurant business shows a cluster of words grouped by each unique type of aspect. For the next step of LDA, based on the outcome, can be distinguishing whether the words in each aspect are positive or negative by doing sentiment analysis. The following figure presents the example of the sentiment analysis result.

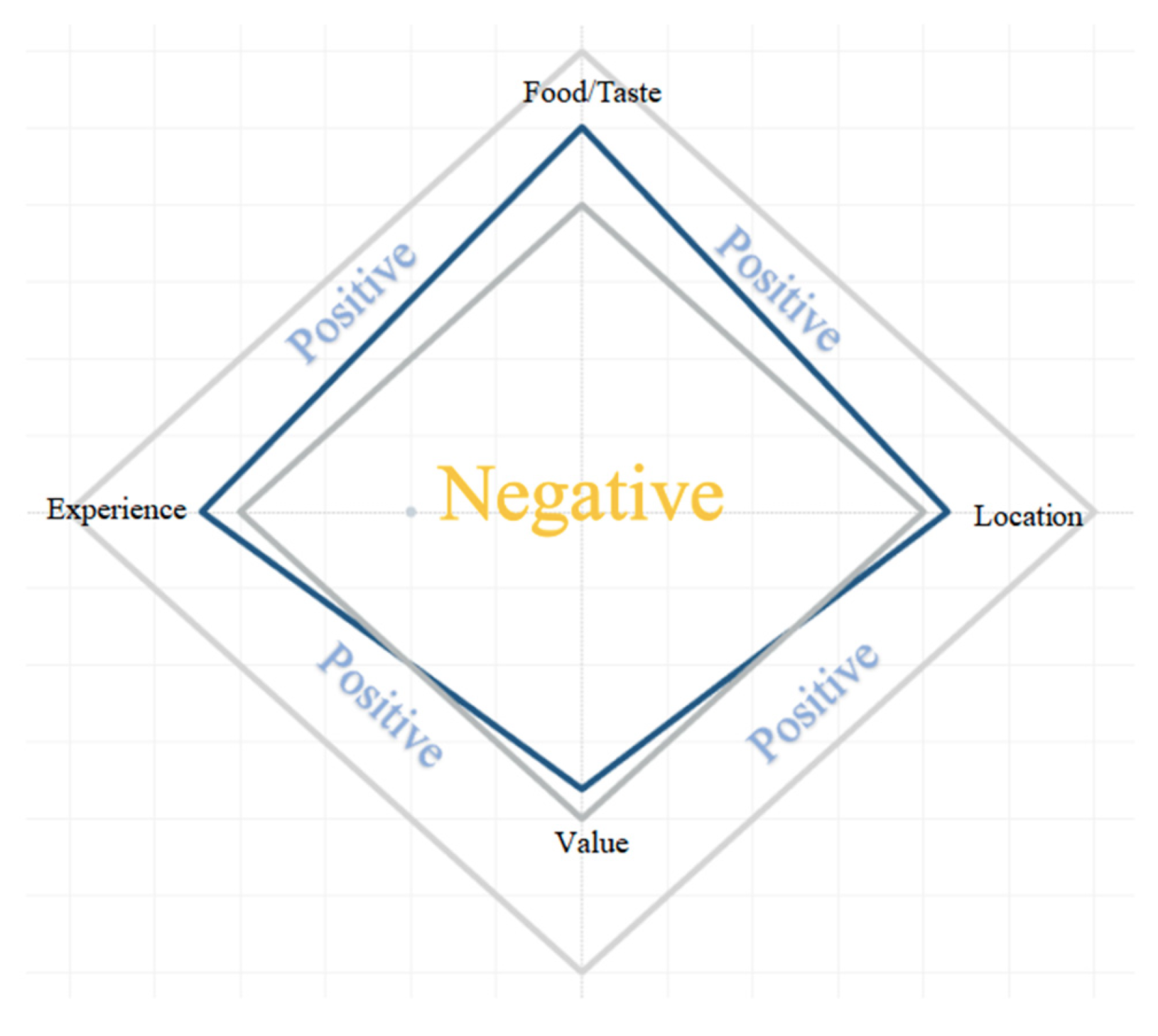


Image 19. Results of Sentiment Analysis of Four Topics.

The sentiment analysis is negative if it points to the inner part of the diamond-shaped rectangle whereas positive sentiment will be outside. Looking at the sentiment analysis results; the business could easily understand which aspect it needs to concern. For example, if a business finds out that one of the aspects is yielding negative sentiment, it can sort by the specific aspect to see the reviews instead of reading all the reviews.

To sum up, we believe that using the LDA method and sentiment analysis will help users and businesses to deeply and specifically understand, and extract the information they need. However, in order to increase the accuracy, the LDA model requires a background industrial knowledge to identify the approximate required number of clusters in each category of the industry. Eventually, we will be able to identify a more specific sorting category of each industry which individual users put more weight on instead of just predicting whether the recent reviews will be potentially useful or not

**8. Conclusion**

Online platforms have made it possible for consumers to gain access to reviews and to provide and consume a wide variety of formats and vast amounts of reviews. Reviews affect the sales of a product or service where it can affect several stages of the consumer's purchasing decision-making process.

This report created a predictive model to classify potential useful and non-useful reviews through Yelp online reviews written from 2006 to 2017. To this end, we created an online review mining with the Naive Bayes algorithm using binary term occurrences vector creation with 94.5% of accuracy. Initially, we defined useful reviews as those that received two or more ‘useful votes. However, after creating training datasets, we could present four categories that identified the features of useful reviews based on the classified results. Due to the special feature of the useful reviews which use capital letters to emphasize users’ experience and comparative or superlative adjectives to describe their experiences, the final model did not need ‘Transform Cases’ and ‘Stemming’ operators which is general text preprocessing requirement.

This report presents features and predictive models of potential useful reviews utilizing Yelp online reviews, and the model has several academic and practical significance; First, the model is expected to help users and businesses to obtain information to take future action by recommending features of potential useful reviews. Second, although the information obtained from the model predicting potential useful reviews is limited, the model proposed a unique feature of useful reviews, and this feature could open a door for further algorithm development by setting more comprehensive classification criteria based on the project’s result. Last but not least, having enough industrial background, conducting LDA analysis could be a powerful tool to improve future models. Specific category sorting systems by Yelp business industry can be expected to generate potential economic effects by improving the quality of customer expectations and experiences as well as enhancing the quality of potential useful review recommendation algorithms.**9. Reference**

* Belt, B. (2017, May 22). *Study: 92% Of Consumers Using Yelp Make A Purchase After Visiting The Platform*. Yelp.

<https://blog.yelp.com/2017/05/study-92-consumers-using-yelp-make-purchase-visiting-platform>.

* Chakure, A. (2020, November 6). *Decision Tree Classification*. Medium. https://medium.com/swlh/decision-tree-classification-de64fc4d5aac.
* Chatterjee, M. (2021, April 19). *The Introduction of KNN Algorithm: What is KNN Algorithm?* GreatLearning Blog: Free Resources what Matters to shape your Career! https://www.mygreatlearning.com/blog/knn-algorithm-introduction/.
* Dato, N. (n.d.). *The Worth of Online Reviews*. Podium. https://www.podium.com/article/online-reviews-worth/.
* Lorenzo, R. (2013, November 4). *What are the differences and similarities between LDA and k-means for topic detection?* Quora. https://www.quora.com/What-are-the-differences-and-similarities-between-LDA-and-k-means-for-topic-detection-assuming-that-I-can-cluster-documents-with-k-means-and-extract-some-common-key-phrases-to-represent-their-topics.
* Luo, Y., Tang, L. (R., Kim, E., & Wang, X. (2020, October 9). *Finding the reviews on yelp that actually matter to me: Innovative approach of improving recommender systems*. International Journal of Hospitality Management. https://www.sciencedirect.com/science/article/pii/S0278431920302498?casa\_token=\_tjwaQxv7eYAAAAA%3Ayna5uB0e6WWsNI77z7o2BNRbhc75DAno9iTOIra\_rHYeCv5Ohj\_3ZeL1n6\_kxlbRJ\_Sk\_NU#bib0065.
* Luo, Y., & Xu, X. (2019, September 25). *Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp*. MDPI. https://www.mdpi.com/2071-1050/11/19/5254/htm.
* Marie, A. (2021, March 11). *Are Yelp reviews reliable?* Online Reputation Management Articles. https://blog.reputationx.com/are-online-reviews-reliable#:~:text=Recent%20research%20suggests%20that%20reviews,positive%20and%20negative)%20are%20fake.
* *Tutorial: Text Mining Using LDA and Network Analysis*. Nodus Labs. (2019, October 8). https://noduslabs.com/cases/tutorial-lda-text-mining-network-analysis/.
* Wiideman, S. (2015, March 15). *Study: How Important are Yelp Reviews Really?* Wiideman Consulting Group. https://www.wiideman.com/blog/local-seo/study-how-important-are-yelp-reviews-really.