Quantitative Analysis for Brand Imaging and Customer Retention

Sonal Sharma
New York University
NY, USA
sonal.sharma@nyu.edu

Karan Gupta
New York University
NY, USA
karan.gupta@nyu.edu

Priyal A. Nile
New York University
NY, USA
pan303@nyu.edu

ABSTRACT

Reviews play an essential role in understanding information about events, places, or any commercial product. These reviews are captured simply on the numeric scale, or it can be elaborated detail text review. In this study, we analyze if we can summarize the tremendous textual criticism by trying to predict ratings on a 5-point scale using review provided by users. We also perform sentiment analysis to understand if the sentiments in the review are following the user star rating provided. Our findings show that star ratings can be predicted to some extent, based on the review comments. The analysis can be leveraged to check the performance of restaurants on different platforms and tracks the performance of the restaurants over the years, which will be beneficial to summarize the review comment to the star rating and save people time to read through the entire review description. Also, it could be used to extrapolate and retain customers from the business perspective of restaurants.

KEYWORDS

Sentiment Analysis, ETL, Classification, Big Data

I. INTRODUCTION

With the advent of the digital world and the evolution of the internet, the rise of the online food ordering system has been gaining significant importance. Adhering to the current market trend by staying adopted to the ever-changing demands of customers, is a challenging task. In our analysis, we have leveraged the data from Yelp, Twitter, and Reddit Dataset.

II. MOTIVATION

Human has always been in search of making human life more accessible and more comfortable. The chain of such comforts and innovative ideas led to the origin of Online Food Ordering. The history of such concept goes back to 1995's, where the first online food ordering service, World Wide Waiter (now known as Waiter.com), was founded in 1995 [8]. The overgrowing hunger of customers and fast delivery systems, which are behind the tremendous success of online food ordering apps and restaurants, have created data as the by-product of such a business. There are no issues when handling a limited number of customers. However, while handling the ever-growing number of users and maintaining their satisfaction led to a new problem of managing the data and processing of such data to sort out all the issues associated with such angry and unhappy customers. We have tried to tackle this issue using the Data processing and analytics technique using the Hadoop Framework projects such as MapReduce and Hive, Impala, HDFS.

This project analyses the reviews posted by crowd-sourced review forum like Yelp, Twitter, and Reddit customers for specific Restaurant Businesses. Polarity (Positive/ Negative/ Neutral) of the review/keywords would be measured using Sentiment Analysis from different forums. Polarity analyzed results would be quantified and would be compared with the actual rating on Yelp, Grub Hub platform.

Relevant keywords taken from the reviews would help in constructing a polarity of the reviews, which would help in deriving insight about the popular perception about a business.

III. RELATED WORK

The study of fellow researchers Michelle Renee D. Ching and Remedios de Dios Bulos, from the Dela Salle University of the Philippines, aims to help restaurants registered in Yelp by recommending business strategies in sustaining and improving their customers' satisfaction through analyzing its customers' text reviews. Although they have used some regression and machine learning techniques for data processing, we are developing our analysis using python and other Hadoop tools like MapReduce, Hive, and Impala [6]

From Princess Sumaya University for Technology of Jordan, the authors Mariam Khader, Arafat Awajan and Ghazi-Al-Naymat have performed a sentiment analysis based on the MapReduce framework. We share some common grounds with their project, such as the usage of MapReduce Framework in our analysis, determining the customer polarity. However, they also have used natural language processing for further classification of the sentiments and reviews based on the machine learning algorithm. But we are not going to incorporate the ML framework in our analysis.[7]

Fellow researcher Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, Jay Ramanathan from the Department of Computer Science and Engineering Ohio State University, has demonstrated the building of Large-scale distributed system for twitter sentiment analysis. They were are concentrating on feeds from twitter only by considering two main components in their study, such as a lexicon builder and a sentiment classifier. Although with the somewhat new name, these two tools are working on the MapReduce framework only. They have performed opinion extraction further through machine learning. We are not incorporating the machine learning framework in our analysis [8]

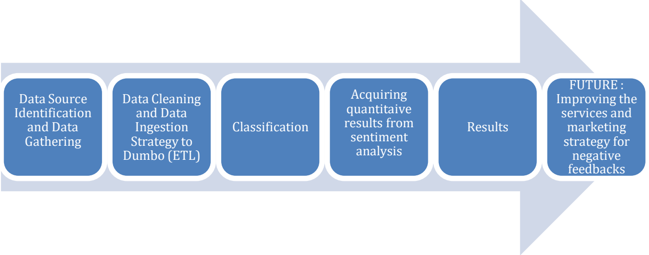
The research paper by fellow authors Kapil Topal and Gultekin Ozsoyoglu Case Western Reserve University are incorporated for the domain Social Network Analysis and Mining. The essence of the project is around the movie ratings and reviews from famous sites such as IMDB or Amazon Prime. Although the idea of this movie review is similar to our project of '*Quantitative Analysis for Brand Imaging and Customer Retention'* in terms of reviewing the polarity and computing the rating based on user reviews, the significant difference is in the mode of analytics that has been incorporated. The Movie Review Analysis is implemented by using the K means clustering machine learning models, whereas our analysis is based on the Big Data Ecosystem [14]

The paper talks about the usage of Big Data technologies to handle a large number of twitter data to perform sentiment analysis on user reviews and determine their polarity. Every year users post a considerable amount of voluminous tweets. In particular, to Big Data, Hadoop technology is used in this paper since it's a scalable open-source framework that can be leveraged to execute operations efficiently on distributed data. Hadoop programming model MapReduce was used for processing and generating big data sets with a parallel, distributed algorithm on a cluster. This research paper helped provide the roadmap to our project 'Quantitative Analysis of Customer Satisfaction and Brand Imaging.' We have taken inspiration from the architectural steps used in this project, which involves careful observation for data sources, data collection, data cleaning, and analysis. The research paper discusses the idea of finding the polarity from reviews which we have also incorporated after reading this paper, and we have also implemented comparing ratings of brands on different platforms which this research paper talks about in its future scope (comparing the reviews of various persons and judge who is the best) [13]

Author Langtao Chen from Missouri University of Science and Technology is assessing the impact analysis of Online Customer Reviews and customer satisfaction based on the evidence collected from Yelp Reviews. It also underlines the churn analysis of customers based on online reviews. This project idea is very much similar to our project idea. However, we are assessing the datasets from Yelp, Reddit, and Twitter, whereas this project is focusing mainly on Yelp Dataset. Also, the author is trying to use some positive and negative words deciding the polarity of their reviews, which is entirely aligned with our analytics. But we are more focused on using Big Data technologies, whereas the author is using the Regression Model of Machine Learning.[15]

IV. DESIGN AND IMPLEMENTATION

We followed the step by step approach as shown in the process diagram to analyze and publish the results of our study


Data Source Identification and Data Gathering
Data Cleaning and Data Ingestion Strategy to Dumbo (ETL)
Classification
Acquiring quantitaive results from sentiment analysis
Results
FUTURE : Improving the services and marketing strategy for negative feedbacks

1. Data Source identification and Data Gathering

We identified Yelp, Twitter, Reddit as our data source, which has a rich dataset of review comments and star ratings. We collected the data for Twitter and Yelp in JSON format, whereas the data for Reddit was collected in CSV format. We have further discussed in datasets in more detail in the DATASETS section.

2. Data Cleaning and Data Ingestion strategy to Dumbo (ETL)

Yelp, Twitter and Reddit data text attribute had characters like @, #, !, ?, spaces. These characters don't add any value in the classification or sentiment analysis process but may add bias in overall analysis, so we cleaned this data by using the process as follows.

- Removed unnecessary spaces after # and @
- Removed null strings
- Dropped empty strings
- Removed @, #,!,?
- Changed the date format to dd-mm-yyyy
- Changed the encoding to ‘utf-8’

Also, the Yelp dataset we have downloaded from their website, which is entirely free, but for the Reddit and Twitter data, we have scrapped from their sites by creating a developer account. Furthermore, WinSCP (Windows Secure Copy), a popular SFTP & FTP Client, was used to copy the files from a local computer to the Dumbo cluster.

All the data that will be used for analysis has been cleaned in this process and can be used for further analysis.

3. Classification

Our research was performed to answer below questions:

1. Can we predict rating based on user’s review?
2. Can we make a model, based on (1) that can be further used for Brand Imaging and Customer Retention?

In our research, the results of question 1 are very critical as it provides a base for correct prediction. Typically, with text limitation on the microblogging site, users tend to put essential and crisp text to convey the message. But Yelp allows users to write free text with no restrictions, unlike Twitter and Reddit, which only allows 140-character limit.

So, it is imperative to evaluate the results of question 1. If reviews resonate with user ratings, then end-users can rely on ratings and do not have to read entire review comments before making any decision. Additionally, reviews can be used to predict Business rating. Based on data analysis, we concluded that stars could have either one of the values from the range [0,5]. Each review text has a star rating assigned to it; hence, text feature is labeled with stars.

4. Acquiring results from Sentiment Analysis

All the cleaned data is stored in tables using Impala, as it provides faster processing. So, now we have all the data in tabular format. Now, we will store the data (present in tables in Impala) into text format, which will be directly stored over Dumbo and can be directly used for analytics (the customer reviews part). Furthermore, we have used TextBlob (an NLTK library). It provides the polarity score, so if the score >0, then the sentiment is positive if score <0, then the sentiment is negative, and if the score =0, then the sentiment is neutral. Furthermore, the polarity score and the user ratings (which are already present in the data) will be used for the comparison for restaurants across different platforms. For testing purposes, we have used ‘Starbucks’ and ‘McDonald's,’ and all the comparisons are made on these only.

V. DATASETS

A. Yelp

The Yelp dataset consists of five data feed, but we primarily work on the below mentioned two feeds.

- Business – Information of all the businesses in Yelp.
- Review – Reviews of all the business.

Yelp provides the data on its website which consists of 85,539 businesses and 2,685,066 reviews.

From this data source, the only following were fetched:

- ID - This contains unique identification for the row data
- Date - The timestamp of a review of the data
- Review - This contains of the text of the review posted by the user
- Business_id - This contains of the business ID of business.
- Stars - This contains of a user star rating for a business.

B. Twitter

All Twitter APIs that return Tweets provide that data encoded using JavaScript Object Notation (JSON). JSON is based on key-value pairs, with named attributes and associated values. These attributes and their state are used to describe objects.

From the data source, the only following were fetched:

- Id - This contains unique identification for the row data
- Created_at - The timestamp of a particular review of the data
- text - This contains the text of the review posted by the user
- location - This contains the business ID of business.

C. Reddit

All the Reddit APIs that return reviews provide that data encoded using Comma Separated Format (CSV).

From this data source, the only following were fetched:

- ID - This contains unique identification for the row data
- Date - The timestamp of a review of the data
- Review - This contains the text of the review posted by the user

VI. RESULTS

I. Comparing the average user ratings with the average polarity ratings

User rating: Rating which user provides while writing the review.

Average polarity ratings: average ratings that we have calculated using the polarity score

We figured the average polarity based on the customer's online reviews out of 5. For this, we scaled the average using the formula average((polarity*5) +5)/2. This made sure to give the average rating out of 5 only.

For Starbucks	
average_user_rating	average_polarity
3.084	2.87

For McDonald's	
average_user_rating	average_polarity
2.125	2.51

1) **For Starbucks:**

% Accuracy for Starbucks = $\frac{100}{1}$ - $\frac{|AveragePolarity - AverageUserRating|}{AverageUserRating} * 100$

= $\frac{100}{1}$ - $\frac{|2.8275-3.084|}{3.084} * 100$

= 100 - 8.317

% Accuracy for Starbucks = 91.68

2) For McDonalds:

% Accuracy for McDonalds = $\frac{100}{1}$ - $\frac{|AveragePolarity - AverageUserRating|}{AverageUserRating} * 100$

= $\frac{100}{1}$ - $\frac{|2.5106-2.125|}{2.125} * 100$

= 100 - 18.145

% Accuracy for McDonalds = 81.85

The model can provide the accuracy between 80-90% and hence can be further used by customers to save the time in reading the reviews. Also, since the model is providing the accuracy between 80-90%, we can set an error margin for +/- .5 for ratings.

II. Comparing the average polarity of the restaurants (Starbucks and McDonalds) across different platforms

For Starbucks:

Over Reddit, average polarity comes out to be 2.62: -->

count	average_polarity
4988	2.62

Over Twitter, average polarity comes out to be 2.68: -->

count	average_polarity
1531	2.68

Over Yelp, average polarity comes out to be 2.83: -->

count	average_polarity
20145	2.83

For McDonalds:

Over Reddit, average polarity comes out to be 2.68: -->

count	average_polarity
4973	2.68

Over Twitter, average polarity comes out to be 2.58: -->

count	average_polarity
2588	2.58

Over Yelp, average polarity comes out to be 2.51-->

count	average_polarity
232	2.51

- For Starbucks, Yelp has got the highest average polarity and it also, suggests that people are more likely to post a review for ‘Starbucks’ on Yelp
- For McDonalds, Reddit has got the highest average polarity and it also, suggests that people are more likely to post a review for ‘McDonalds’ on Reddit.

III. Comparing the average user ratings and average polarity ratings yearly

For Starbucks:

period	average_user_rating	average_polarity
2006	2.25	2.04
2007	3.7	2.66
2008	3.61	2.93
2009	3.36	2.88
2010	3.52	2.95
2011	3.43	2.95
2012	3.47	2.97
2013	3.45	2.95
2014	3.27	2.89
2015	3.14	2.84
2016	3	2.81
2017	2.89	2.77
2018	2.97	2.78
2019	2.96	2.77
2020	3.07	2.82

For McDonalds:

period	average_user_rating	average_polarity
2010	3.5	2.84
2011	2	2.83
2012	3	2.62
2013	2.78	2.46
2014	1.65	2.35
2015	2.4	2.55
2016	2.33	2.54
2017	1.92	2.51
2018	2.02	2.53
2019	1.95	2.48
2020	1	2.03

- Seeing the above results, it's clear that Starbucks does the great business between the year 2010-2014, and after that, their ratings declined between the year 2016-2019.
- McDonald's average ratings declined between the years 2010-2014, but it shows progress in 2015 and after.

IV. *Comparing the user ratings and average polarity ratings for different regions*

Starbucks:

State	average_user_rating	average_polarity
WI	3.13	2.84
ON	3.4	2.94
NV	2.9	2.77
SC	3.23	2.81
QC	3.64	2.94
NC	3.08	2.82
OH	3.43	2.88
IL	3	2.9
AZ	3.04	2.82
PA	3.45	2.94
AB	3.23	2.91

McDonalds:

State	average_user_rating	average_polarity
ON	2.36	2.55
NV	2.38	2.75
NC	1.78	2.6
OH	1	1.89
IL	3.2	2.74
AZ	1.52	2.32
AB	1.83	2.38

- From the above results, we can see that in the state 'QC' average user rating is 3.64, and the average polarity rating is 2.94, which is the highest among all. So, we can conclude that the STARBUCKS being the most popular brand in state 'QC' among all other states.
- From the above result, we can see that in the state, 'IL' average user rating is 3.18, and the average polarity rating is 2.74, which is the highest among all for McDonald's. So, we can conclude that McDonald's is the most popular brand in state 'IL' among all states.

V. *Seeing which items are influencing the sales of restaurant:*

Starbucks:

Features	Frequency
drive thru'	2948
customer service	1868
every time	802

starbucks location	720
parking lot	619
iced coffee	560
green tea	537
worst starbucks	478
don’t know	447

McDonalds:

Features	Frequency
ice cream	74
delicious ice	46
drive thru	42
mcdonalds ceo	30
outta hand	25
advertising getting	25
mcdonalds advertising	25
getting outta	25
lol that’s	24
cream machine	19
McDonalds sprite	19
McDonalds money	18

From the above results we can conclude that:→

- For Starbucks, these are some of the items which people were frequently talking about:→ drive thru, iced coffee, customer service
- For McDonalds, there are some of the items which people are talking about: ice cream, drive thru, advertising.

This basically helps in understanding the items where they are really doing good (see the word count) and can helps them in improving the areas where they are falling behind to retain the customers for their business.

VII. Future Work

We want to incorporate various aspects like the ambiance, business, cleanliness, desserts being offered, location, food, payment, and reservation, etc. in our analytic. Based on the sentiments and feedbacks received from the customer, these analytics would further try to suggest the best possible actions which could curb these negative reviews also.

VIII. Conclusion

Big Data ecosystem underlines the importance of parallel & distributed computing systems with an optimum time of execution. In this study, we have reviewed Yelp, Twitter, and Reddit data to understand if we can create a quantitative model out of the reviews that could help in understanding the customer reviews in an efficient way. With this said, it may help the restaurant in understanding their products and brand image among users and understand which business strategy they are going good or bad.

We are also able to come up with the comparisons that could be made across the regions, timeline, and review platforms for different restaurants, and that gives the much better picture to the restaurants on parts where they are doing good business. Furthermore, it may also help them in coming up with a strategy for customer retention in areas where they are lagging. Again, the whole analysis is based on the fixed set of datasets, so the results may likely change over time, as the data size increases, but with the use of Big data, we can achieve the computations in very less time. With this project, we are also able to appreciate the power of Hadoop technology and its ability to handle large datasets in seconds..

IX. Acknowledgment

There are several resources which are playing a vital role in the development of this project. To name a few, Professor Suzanne McIntosh, grader Omkar Patinge & Srishti Grover, and NYU HPC Team. We would also like to thank Yelp Team, Twitter Team, and Reddit for allowing us to use their data for analysis.

X. References

[1] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In Proceeding of the 18th ACM conference on Information and knowledge management, pages 375–384, New York, NY, USA, 2009. ACM.

[2] Gayatree Ganu, NoAľmie Elhadad, and AmAľlie Marian. [n. d.]. Beyond the Stars: Improving Rating Predictions using Review Text Content. ([n. d.]).

[3] Himanshu Lohiya. 2018. Sentiment Analysis with AFINN Lexicon. Web Page. (July 2018). https://medium.com/@himanshu_23732/ sentiment-analysis-with-afinn-lexicon-930533dfe75b

[4] Isa Maks and Piek T. J. M. Vossen. 2013. Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions? In RANLP

[5] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

[6] Michelle Renee and Remedios de,2019 “Improving Restaurants Business Performance Using Yelp DataSets through Sentiment Analysis” https://dl-acm org.proxy.library.nyu.edu/citation.cfm?id=3340018

[7] Mariam K, Arafat A, Ghazi A,2018 " Sentiment Analysis Based on Map Reduce : A survey"https://dl-acm-org.proxy.library.nyu.edu/citation.cfm?id=3291795
[8] Vinh N,Chaitanya S, Rajiv R, Jay R,2012 "Towards Building large-scale Distributed system for twitter-sentiment analysis " https://dl-acm-org.proxy.library.nyu.edu/citation.cfm?id=2245364
[9] TextBlob. 2019. TextBlob. web. (2019). https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis
[10] Woo, Jongwook., 2013. "Market Basket Analysis Algorithms with MapReduce" DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28, 2013, Volume 3, Issue 6, pp. 445-452, ISSN 1942-4795.
[11] Yelp. 2019. Yelp Dataset JSON. web. (2019). https://www.yelp.com/dataset/ documentation/main
[12] Boya Yu, Jiaxu Zhou, Yi Zhang, and Yunong Cao. 2017. Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. (09 2017)
[13] Anusha N, Divya G, Ramya B. (2017)Sentiment analysis for twitter data through Big Data : https://www.ijert.org/research/sentiment-analysis-of-twitter-data-through-big-data-IJERTV6IS060183.pdf
[14] Kamil Topal and Gultekin Ozsoyoglu (2016) https://dl.acm.org/citation.cfm?id=3192641
[15]Langtao Chen(2019- The Impact of the Content of Online Customer Reviews on Customer Satisfaction: Evidence from Yelp Reviews https://dl.acm.org/citation.cfm?id=3359448
[16] H. M. Wallach. Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning, pages 977–984, New York, NY, USA, 2006. ACM.
[17] Y. Oh, S. Chae, "Movie Rating Inference by Construction of Movie Sentiment Sentence using Movie comments and ratings", Journal of Internet Computing and Services (JICS) 2015. Apr.: 16(2): 41-48.
[18] J. Jo, S. choi, "Sentiment Analysis of movie review for predicting movie rating", damis, 34(3) 2015.09, 161-177.
[19] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In CIKM '06, pages 51–57, New York, NY, USA, 2006. ACM.