International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I12P102

Volume 71 Issue 12, 5-11, December 2023
© 2023 Seventh Sense Research Group®

Original Article

Sentiment Analysis for Online Reviews for Brand Imaging and Customer Retention

Karan Gupta[1], Amit Bhanushali[2]

[1]*Sun Power Corporation, Austin, TX.*
[2]*Department Independent Researcher, West Virginia University, WV, USA.*

[1]*Corresponding Author : karangupta485@gmail.com*

Received: 10 October 2023 Revised: 14 November 2023 Accepted: 01 December 2023 Published: 15 December 2023

Abstract *- Reviews play an essential role in understanding information about events, places, or any commercial product. These reviews are captured simply on a numeric scale, or they can be elaborated in detailed text reviews. In this study, we analyze if we can summarize the tremendous textual criticism by trying to predict ratings on a 5-point scale using reviews 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 track 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.*

1. 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 adapted 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.

2. Motivation

Humans have 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 a concept goes back to 1995 when the first online food ordering service, Worldwide 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, handling the ever-growing number of users and maintaining their satisfaction led to a new problem of managing the data and processing 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, and HDFS.

This project analyses the reviews posted by crowd-sourced review forums like Yelp, Twitter, and Reddit customers for specific Restaurant Businesses. The 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 and Grub Hub platform.

Relevant keywords taken from the reviews would help construct a polarity of the reviews, which would help derive insight into the popular perception of a business.

3. 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 with Yelp by recommending business strategies for 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]

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

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 using MapReduce Framework in our analysis to determine customer polarity. However, they also have used natural language processing to further classify the sentiments and reviews based on the machine learning algorithm. However, we are not going to incorporate the ML framework in our analysis.[7]

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

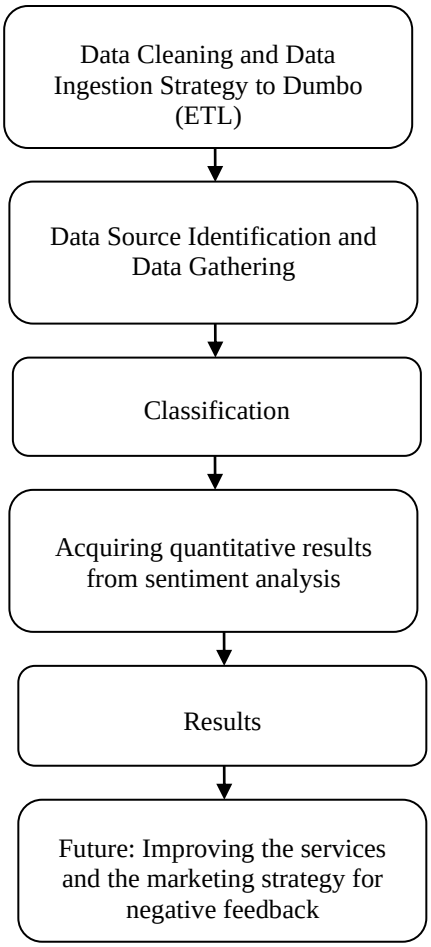
The research paper by fellow authors Kapil Topal and Gultekin Ozsoyoglu Case Western Reserve University is incorporated into the domain of Social Network Analysis and Mining. The essence of the project is around movie ratings and reviews from famous sites such as IMDB or Amazon Prime. Although the idea 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 using 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, for Big Data, Hadoop technology is used in this paper since it is 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 work 'Sentiment Analysis for Online Restaurant Reviews for Brang Imaging and Customer Satisfaction.' We have taken inspiration from the architectural steps used in this project, which involves careful observation of 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 focuses mainly on Yelp Dataset. Also, the author is trying to use some positive and negative words to decide the polarity of their reviews, which is entirely aligned with our analytics. However, we are more focused on using Big Data technologies, whereas the author uses the Regression Model of Machine Learning.[15]

4. 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.



4.1. Data Source Identification and Data Gathering

We identified Yelp, Twitter, and Reddit as our data sources, which have 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 datasets in more detail in the DATASETS section.

4.2. Data Cleaning and Data Ingestion Strategy to Dumbo (ETL)

Yelp, Twitter, and Reddit data text attributes had characters like @, #, !, ? and spaces. These characters do not add any value in the classification or sentiment analysis process but may add bias in the 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.

4.3. Classification

Our research was performed to answer the following:

1. Can we predict the rating based on the 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 they provide a base for correct prediction. Typically, with text limitations on the microblogging site, users tend to put essential and crisp text to convey the message. But Yelp allows users to write free text without restrictions, unlike Twitter and Reddit, which only allow a 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 ratings. 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, the text feature is labeled with stars.

4.4. 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) in 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 the 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 of restaurants across different platforms. For testing purposes, we have used 'Starbucks' and 'McDonald's,' and all the comparisons are made on these only.

4.5. Datasets

4.5.1. Yelp

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

- 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 a 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.
- Business_id - This contains the business ID of the business.
- Stars - This contains a user star rating for a business.

4.5.2. 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 a 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 the business.

4.5.3. 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 a 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.

5. Results

5.1. Comparing the Average User Ratings with the Average Polarity Ratings

5.1.1. User Rating

Rating that the user provides while writing the review.

5.1.2. 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

5.1.3. For Starbucks

% Accuracy for Starbucks = 100 – 100* (|Average Polarity – AverageUserRating| / AverageUserRating)

= 100- 100*(|2.2875 – 3.084| / 3.084)
= 100 – 8.317
= 91.68
% Accuracy for Starbucks = 91.68

5.1.4. For McDonalds

% Accuracy for McDonalds = 100 – 100* (|Average Polarity – AverageUserRating| / AverageUserRating)

= 100- 100*(|2.5106 – 2.125| / 2.125)
= 100 – 18.145
= 81.85
% Accuracy for McDonalds = 81.85

The model can provide accuracy between 80-90% and hence can be further used by customers to save time reading the reviews.

Also, since the model provides an accuracy between 80-90%, we can set an error margin for +/- .5 for ratings.

5.2. Comparing the Average Polarity of the Restaurants (Starbucks and McDonalds) Across Different Platforms

5.2.1. For Starbucks

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

count	average_polarity
4988	2.62

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

count	average_polarity
1531	2.68

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

count	average_polarity
20145	2.83

5.2.2. For McDonalds

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

count	average_polarity
4973	2.68

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

count	average_polarity
2588	2.58

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

count	average_polarity
232	2.51

- For Starbucks, Yelp has 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.

5.3. Comparing the Average User Ratings and Average Polarity Ratings Yearly

5.3.1. 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

5.3.2. 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 is clear that Starbucks did 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 and 2014, but it showed progress in 2015 and after.

5.4. Comparing the User Ratings and Average Polarity Ratings for Different Regions

5.4.1. 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

5.4.2. 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’, the 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 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’, the 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.

5.5. Seeing which Items are Influencing the Sales of Restaurants

5.5.1. 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

5.5.2. 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, and customer service.
- For McDonalds, there are some of the items which people are talking about: ice cream, drive-thru, and advertising.

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

6. Future Work

In this analysis, we utilized restaurant review data only until the year 2019 due to availability constraints. An important area for future work is to expand the analysis to more recent data from 2020 onwards. Adding reviews from the last 2-3 years could provide valuable insights into changing customer sentiment, particularly in light of events like the COVID-19 pandemic that have significantly impacted the restaurant industry. Analyzing reviews during 2020-2022 and comparing them to prior years could reveal interesting trends and shifts in brand perception. Furthermore, incorporating demographics and semantic analysis of review aspects could enrich the insights extracted from more recent data. Overall, expanding the data timeframe studied and enhancing the analytical techniques on newer data represents a logical next step for this research.

7. Conclusion

The Big Data ecosystem underlines the importance of parallel and distributed computing systems with an optimum execution time. 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 us understand customer reviews in an efficient way. With this said, it may help the restaurant understand 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 comparisons that could be made across the regions, timelines, and review platforms for different restaurants, and that gives a much better picture of 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 a 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 much less time. With this work, we can also appreciate the power of Hadoop technology and its ability to handle large datasets in seconds.

Acknowledgment

There are several resources which are playing a vital role in the development of this project. We would also like to thank the Yelp Team, Twitter Team, and Reddit for allowing us to use their data for analysis.

References

[1] Chenghua Lin, and Yulan He, "Joint Sentiment/Topic Model for Sentiment Analysis," *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, pp. 375-384, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[2] Gayatree Ganu, Noemie Elhadad, and Amelie Marian, "Beyond the Stars: Improving Rating Predictions using Review Text Content," *12th International Workshop on the Web and Databases (WebDB 2009), Providence*, Rhode Island, USA, pp. 375-384, 2009. [Google Scholar] [Publisher Link]

[3] Himanshu Lohiya, Sentiment Analysis with AFINN Lexicon, 2018. [Online]. Available: https://himanshulohiya.medium.com/sentiment-analysis-with-afinn-lexicon-930533dfe75b

[4] Isa Maks, and Piek Vossen, "Sentiment Analysis of Reviews: Should we Analyze Writer Intentions or Reader Perceptions?," *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 415-419, 2013. [Google Scholar] [Publisher Link]

[5] Data Dictionary: Standard V1.1, Developer Platform. [Online]. Available: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

[6] Michelle Renee D. Ching, and Remedios de Dios Bulos, "Improving Restaurants Business Performance Using Yelp Datasets through Sentiment Analysis," *Proceedings of the 3rd International Conference on E-commerce, E-Business and E-Government, ACM*, pp. 62-67, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mariam Khader, Arafat Awajan, and Ghazi Al-Naymat, "Sentiment Analysis Based on Map Reduce: A Survey," *Proceedings of the 10th International Conference on Advances in Information Technology, ACM*, PP. 1-8, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Vinh Ngoc Khuc et al., "Towards Building Large-Scale Distributed System for Twitter-Sentiment Analysis," *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 459-464, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sentiment Analysis, TextBlob, 2019. [Online]. Available: https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis

[10] Jongwook Woo, "Market Basket Analysis Algorithms with MapReduce" *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 445-452, 2013.[CrossRef] [Google Scholar] [Publisher Link]

[11] Yelp Open Dataset, Yelp Dataset, 2019. [Online]. Available: https://www.yelp.com/dataset/

[12] Boya Yu et al., "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews," *Computation and Language, Cornell University*, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] N. Anusha, G. Divya, and B. Ramya, "Sentiment Analysis for Twitter Data Through Big Data," *International Journal of Engineering Research & Technology*, vol. 6, no. 6, pp. 307-309, 2017. [Google Scholar] [Publisher Link]

[14] Kamil Topal, and Gultekin Ozsoyoglu, "Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1170-1176, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Langtao Chen, "The Impact of the Content of Online Customer Reviews on Customer Satisfaction: Evidence from Yelp Reviews," *Companion: Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, pp. 171-174, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Hanna M. Wallach, "Topic Modeling: Beyond Bag-of-Words," *Proceedings of the 23rd International Conference on Machine Learning, ACM*, pp. 977-984, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yean-Ju Oh, and Soo-Hoan Chae, "Movie Rating Inference by Construction of Movie Sentiment Sentence using Movie Comments and Ratings," *Journal of Internet Computing and Services*, vol. 16, no. 2, pp. 41-48, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[18] Jo Jung-Tae, and Choi Sang-Hyun, "Sentiment Analysis of Movie Review for Predicting Movie Rating," *Management and Information Systems Review*, vol. 34, no. 3, pp. 161-177, 2015. [Google Scholar] [Publisher Link]

[19] Z. Zhang, and Balaji Varadarajan, "Utility Scoring of Product Reviews," *Proceedings of the 15th ACM international conference on Information and Knowledge Management*, pp. 51-57, 2006. [CrossRef] [Google Scholar] [Publisher Link]