Big Data Analytics Symposium - Fall 2019

Analytics Project: Quantitative Analysis for Brand Imaging and Customer Retention

Team Name : Zerobug

Team: Sonal Sharma, Karan Gupta & Priyal A. Nile

Abstract: Reviews of events, places, or any commercial product play an essential role in influencing the decision making of business. These reviews are captured simply on the numeric scale, or it can be elaborated detail text review. In this study, we analyzed the voluminous textual information to predict ratings on a 5-point scale using reviews provided by users. We have also performed sentiment analysis to understand if the sentiments in the review are following the user star rating provided. Our findings also highlight most talked things about a restaurant in the user reviews.

Motivation

Who are the users of this analytic? Stakeholders, Customers, Competitors

Who will benefit from this analytic? Sellers and Service Providers

Why is this analytic important?

The analysis can be leveraged to analyze restaurants image on different platforms, time periods, demographics and verify the user provided ratings with ratings quantified from text reviews. Additionally, it saves time of end users since they don't need to read through the entire review set for finding about the most talked about things in reviews. Also, key insights from this analytic can be used to extrapolate to customer retention.

Goodness

What steps were taken to assess the 'goodness' of the analytic?

We have compared the results and verified the predicted ratings with a crowd-sourced review forum for a restaurant chain. Verifying the results with crowd-sourced forums would be best since it represents the current scenario of a brands image through reviews. We have considered some error margin in our model since the restaurant ratings at crowd-sourced forum are based on average user ratings whereas our model would generate content-based ratings using Sentiment Analysis.

Data Sources

Name: Yelp Open Dataset (https://www.yelp.com/dataset)

Description: This Yelp dataset is a subset of Yelp's businesses, reviews, and user data for use in personal, educational, and academic purposes.

Size of data: 8 GB

Name: Twitter

Description: Twitter contains a lot of useful information, that could be useful for us in finding the information about restaurants.

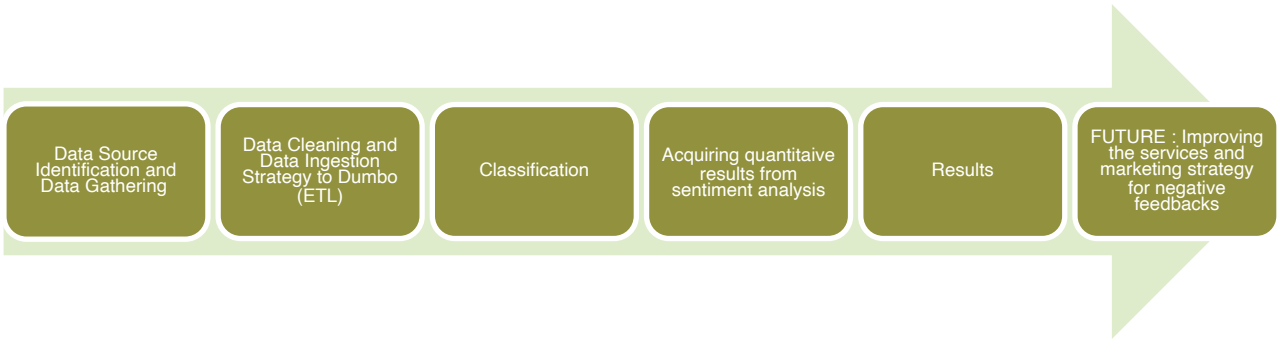
Size of data: 1.2 GB

Name: Reddit

Description: Reddit contains a lot of useful information, that could be useful for us in finding the information about restaurants.

Size of data: 1GB

Design Diagram



Platform(s) on which the analytic ran:

NYU Dumbo Cluster.

Results

Factors	*Starbucks*	*McDonald's*
Average User Ratings	*3.08*	*2.125*
Average Polarity		
On Yelp	*2.83*	*2.51*
On Twitter	*2.67*	*2.58*
On Reddit	*2.62*	*2.67*
Total Average Polarity	*2.82*	*2.51*
Accuracy	*91.5%*	*81.8%*

Average Polarity Based on Locations	*Starbucks*	*McDonald's*
ON	*2.94*	*2.55*
NV	*2.77*	*2.75*
SC	*2.77*	*2.6*
Average Polarity Based on Time Period		
2017	*2.77*	*2.51*
2018	*2.78*	*2.53*
2019	*2.77*	*2.48*

Most Talked about factors			
Starbucks	*Count*	***McDonald's***	*Count*
drive thru'	*2948*	*ice cream*	*74*
customer service	*1868*	*delicious ice*	*46*
parking lot	*619*	*drive thru*	*42*
worst starbucks	*478*	*advertising getting*	*25*

Obstacles

OBSTACLE-1: We have used NLTK libraries like TextBlob, stop words, ngrams but these were not present on the node. Unaware of this, our code was getting failed every time. Later, Wensheng from HPC team cooperated with us in getting the NLTK libraries installed over the node. Though the libraries were later installed, but we could not easily run it over the node as a python wrapper script was needed to run it. Thus, we created this additional python script for execution of mapper code in the pipeline.

OBSTACLE-2: MapReduce jobs were failing in fetching the data from hive tables to hdfs. TA's helped us in figuring out that it's best to create an external table instead of manage table since external tables can access data stored in sources like remote HDFS locations.

Quantitative Analysis for Brand Imaging and Customer Retention

Summary

In this study, we have reviewed Yelp, Twitter, and Reddit data to create a quantitative model based on the customer review data, which will help in understanding a brands image in an efficient way.

Analyzed results can be interpreted as comparative analysis of user reviews on different social platforms of a brand during a timeline across the regions.

Furthermore, this model will also help the end users(business) in producing a strategy for *customer retention* and improving *brand image*.

The whole analysis is based on fixed set dataset, so the results may likely change over time and based upon data volume. Our model is capable enough to handle large dynamic datasets because of Hadoop Big data technology, which helps achieving the computations in very less time.

Acknowledgements

We would like to thank NYU HPC team for there continuous support. We would also like to thank Prof. Suzanne Mcintosh and TA's Omkar and Srishti for helping us out with our doubts. Last but not the least we would also like thank Yelp, Twitter and Reddit for allowing us to use their data for our analysis.

Quantitative Analysis for Brand Imaging and Customer Retention

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, NoAl'mie Elhadad, and AmAl'lie 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] 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

[6] 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

[7] 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

Thank you!