Developing techniques for finding the correlation between user reviews and userratings

BY

MD. ABDULLAH AL MAMUN
ID: 152-15-5790

AND

MD. SAKIBUL HASSAN
ID: 152-15-5782

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. SAMIA NAWSHIN
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY, 2019

APPROVAL

This Project titled **Developing techniques for finding the correlation between user reviews and user ratings**, submitted by Md. Abdullah Al Mamun, ID No: 152-15-5790 and Md. Sakibul Hassan, ID No: 152-15-5782 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 04-05-2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Tarek Habib — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Moushumi Zaman Bonny — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Saddam Hossain — **External Examiner**
Assistant Professor
Department of Computer Science and Engineering
United International University

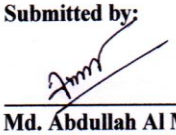
DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms. Samia Nawshin, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

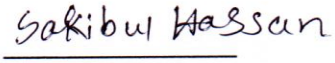
Supervised by:

Ms. Samia Nawshin
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md. Abdullah Al Mamun
ID: 152-15-5790
Department of CSE
Daffodil International University



Md. Sakibul Hassan
ID: 152-15-5782
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

FirstWe express our heartiest thanks and gratefulness to almighty Allah for His divine blessing makes it possible to complete the final year thesissuccessfully.

We are really grateful and wish our profound indebtedness to **Ms.SamiaNawshin**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of oursupervisor in the field of *Data Mining*helped us a lot to carry out this research. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice,reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Almighty Allah and**Dr. Syed Akhter Hossain, Professor & Head,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staffs of CSE department of Daffodil International University.

We would like to thank **Mr. Mehedi Imam Shafi**, R&D Engineer, CodeMarshal, for his enormous help, support and direct guidance throughout the research to enable us reaching the stage that our work currently is in.

We would like to thank our entire course mates in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant supports and patients of our parents.

ABSTRACT

In order to predict how a user will respond to a product, we mustuncover the tastes of the user and the properties of the product. Forexample, in order to predict whether a user will enjoy a food or not it depends on theuser's level of interest in which type of foods user likes most. User feedback is required todiscover these dimensions, which comes in the form of ratings andreviews. In this thesis, we aim to find a correlation between user reviews and star ratings. However, traditional methods discard review text, whichmakes these latent factors difficult to interpret.In this thesis, Firstly, our approach is to analyze a text review according to the texts and find the sentiment of the user review whether it is a positive review or negative review. Secondly, after getting the sentimental analysis,finding the correlation between the sentimental analysis and user star ratings. our approachmore accurately predicts sentiments by analysing the informationpresent in review text. Our discovered methodology can be used to identify the useful and representative reviews.

Table of Contents

List of Figures

FIGURES **PAGENO**

List of Tables

Tables **PAGE NO**

CHAPTER 1
Introduction

1.1 Introduction

User generated reviews have been a great influence in decision making nowadays. Usually a public business runs on the basis of public reactions. People take decisions on the basis of other user's reviews. Asan example, if someone wants to go to a restaurant, he or she will search for the positive reviews about some restaurants, then the right place will be decided. So, user's review has a large impact on decision making. Even, an owner can understand the state of his restaurant by reading the user's reviews. But in most of the cases it is difficult and time consuming to read a large amount of text reviews. In this situation, a visualization that summarizes the user generated review can help to understand the state or taking decisions. Our model has the potential to process usergenerated reviews and provide more context than only quantitativeinformation.

Our model visualizes the summary of the correlation between user generated reviews and star ratings they give. But the generated review is analyzed whether, the review is positive or negative. The correlation implies that the user generated review and rating are linearly correlated or not. This visualization or summarization will help customers or the business holders to take decision according to the analysis.

1.2 Motivation

Sentiment analysis on big data using data mining algorithms has always been an interesting area of research. The reasons are many in numbers. As, people depends on the user generated text reviews to take a decision, so analyzing the user generated review text data and make a summarization is an important factor. Finding a correlation between the user reviews and star ratings can visualize or summaries the whole scenario and it will be easier to understand the sentiment. The correlation coefficient value implies that the reviews and ratings are positively correlated or negatively correlated.

There aremany existinganalyses on user generated text reviewbut there are no findings of correlation between text review and star rating. This analysis could be a step ahead to analyzing text reviews generated by users.

1.3 Rationale of the Study

Data mining is a fairly old area of research. This problem is almost a new dimensional problem as the finding of correlation.The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables [1]. This statistical measure can imply that the review texts generated by the user are positively correlated or negatively correlated. Since, there is a statistical measure so there will be an understanding of positive and negative sentiments of users. Every work done in the past (except a few) have been done on the text mining and analyzing and sentiment analysis. Therefore, the problem still exists as to work from the very fundamental. Researches is still going on to solve this problem moreefficiently.

1.4 Research Questions

Every model is basically made up of many small simple parts. As such a research problem often becomes understandable once every part of the problem is defined properly. The research this report is made upon also comes with many sub problems that needs to be understood before understanding the actual problem. This section of the report deals with all the basic questions that comes in with the topic itself.

1.4.1 User Generated Text Review

A text review is a string produced by a user using user's native language. This includes all structured and unstructured sentences. The sentence may or may not follow rules of grammar. The sentence even may or may not a complete sentence. For human being this is very normal to converse in natural language. The language (any one language) is being developed for thousands of years and thus now very complex.

1.4.2 Analyzing Text Review

Our dataset contains user generated text reviews which are natural languages. To analyze the reviews, we had to process the data set. As the data processing has been done by the theory of NLP (Natural Language Processing) [2]. After processing the data, we analyzed the data to get the sentiment of people which was generated by them as a text review. The sentimental analysis was made up as polarity and subjectivity. Polarity describes, the review is positive or negative and the Subjectivity implies that the review is fairly subjective or not.

1.4.3 Finding Correlation

The goal of this research is to propose, a methodology which can find the correlation between user generated text review and star ratings. The user generated text is analyzed as, the review is positive or negative and the review is fairly subjective or not. Then the value of polarity and star ratings have been used to find the correlation coefficient value using PCC (Pearson Correlation Coefficient) Parametric methods [3].

This research is limited to work with the natural language English for now. The reason of choosing English is, English is an international and vastly used language all over the world and the data set we collected was in English.

1.5 Expected Outcome

This research includes both a proposal to solve the stated problems and a basic implementation to test the hypothesis. The primary objective of this project is to propose a methodology which can find the correlation between user generated text review and star rating. That is to propose a sentiment data analysis for the problem. This problem includes many sub objectives. Among the sub objectives it is expected to solve the following ones -

- Analyzing the sentiment of the text review.
- Ability to detect the review whether it is positively explained or negatively explained.
- Detecting the subjectivity of the text review.
- Finding the correlation coefficient between text review and star rating.

1.6 Report Layout

This report is designed in such a way that, it will give the proper idea of the problem and the findings as well as the research methodology. The report is followed by the standard thesis reporting template provided by DIU.

First chapter gives the summary of the problem, objective of the research,motivation and the introduction of the report.

Chapter 2contains the background and the related works. There is a brief discussion of related works, scope of the problems as well as research summary. Even we have described the challenges we had during the work.

Chapter 3discusses the research methodology and techniques we used in the research in details. As well as the data collection process and implementation requirements. This chapter will give the whole idea of our approach to solve the problem.

Chapter 4 contains the details of implementation. The proposed methodology has been implemented into a python script which has been described in detail using the proper diagram and contents.

Chapter 5 deals with the finding of the research and the result we got. Accuracy of the findings and detailed description with statistical analysis.

The final chapter containsthe whole summary of our working approach, conclusion, some limitations and future scope as well.

CHAPTER 2
Background

2.1 Introduction

Sentiment Analysis is the process of identifying the pattern of information and categorizing the data computationally from a piece of text and determining the data whether it is positive, negative or neutral. We all know the success of company or product directly depend on the customer choice. If the customer likes the product then it is considered as a success but if not, then the company certainly need analyzing the situation where data mining comes on. Data mining is very much important to determine the opinion of a customer about a specific business.

Like we said earlier, our research goal is to make a system which can detect text emotions and make a comparison of the attributes of information. It is easy to excess large number of different kinds of built in library in AI which has been developed by different types of programming language.

2.2 Related Works

Now-a-days online reviews are very much important for the customer to make a good decision of the product or seeing a movie, going around a place, going to a restaurant. These online reviews are the information for sentiment analysis. It is easy to classify the information as it has three aspects as positive, negative and neutral.

There have been many researches on this specific problem. Both proposals and systems have been made to solve the problem efficiently. For the last few years a lot of new approaches and ideas have been introduced to analyze the data.Everywork has its pros and cons. All proposals, made, have been helpful to the pathway for a better solution later on.

Xing Fan[4] has developed a system using reviews data to identify a pattern of negation phrases. Sentence level and review level classification of data performed for the data collected from Amazon reviews in February to April 2014.

Aashutosh Bhatt [5] used reviews of iphone5 scrapped from Amazon website and suggested rule-based extraction of product feature sentiment analysis. POS method are used to implement each and every sentence level and the results are shown in charts.

Cane-Wing [6]has developed a system to elicit user preferences expressed in textual reviews. Mapping such preferences onto some rating scales that can be understood by existing CF algorithm. One important task in their rating inference framework is the determination of sentimental orientations (SO) and strengths of opinion words.

Theresa Wilson [7]has presented a system phrase-level sentiment analysis that first determines the whether an expression is neutral or polar then disambiguates the polarity of the polar expressions. With this approach, the system is able to identify the large set of expression of sentiment analysis.

Sasikala[8] dealt with the reviews of the customer where the system predicted the sentiment analysis with empty or blank rating using various sentiment analysis classifiers. The results of the data compared and classified the review text into positive, negative or neutral.

All the research that have been done for data mining is the way of opening new ideas. It is possible to think innovative in this field. Predicting data, finding the pattern of unexpected thing will be a matter of interesting for the people. People has their own way of working style, so as we.

All the research or system that we mentioned here helped us to think differently in order to reach a good solution.

2.3 Research Summary

We have analyzed the textreviews generated by user and all the rating stars to make a very clear understanding of the problem. The research included some of other authors' works partiallytoget better ideas, mathematical representation of othersworks. The research shows that the correlation still needs to be accurate to representa better outcome. Our research and proposed work have been tested and proved tobe a working materialfor further implementation. Our system can generate better accuracy for any given data for any attached reviews. Although our works also have its own limitations and challenges and also some future scopes of work which will be discussedlater with details.

2.4 Scope of the Problem

The working scope of our project will be covering a large number of problems. The future of sentiment analysis is going to go big, and faster. Along with the reviews now-a-days sentiment analysis in social media creates a long effect in terms of determining user behavior. The surface of the like, comments, and shares aim to reach and truly understand the significant meaning of interacting with the data.As a result of deeper and better understanding of the feelings, emotions and sentiments of a brand or organization's key, high-value audiences, members of these audiences will increasingly receive experiences and messages that are personalized and directly related to their wants and needs. Rather than segment markets based on age, gender, income and other surface demographics, organizations can further segment based on how their audience members actually feel about the brand or how they use social media. While some people shudder at the thought of companies learning more about them, more exact targeting means that, in the near future, we will no longer be scratching our head wondering why we see advertisements for products we'd never dream of purchasing

In our project we used our training dataset containing user reviews and star ratings to represent a relation. As our data is very new and no work had not been done yet, so it is a great opportunity to go deeper with the research, and finding interesting pattern with the data set.

2.5 Challenges

The problem is not easy to deal with, there are lot of small problem that needs to be solved first. We needed reviews along with star ratings information and also checked them whether the information was valid or not. Feature selection, word tokenization and word selection were so much important for this research. Converting the star ratings into numeric value was a critical task for us, and also chose the perfect algorithm and library to work with.

There were lot of similar words in our dataset which needed to identify and cleaned also. Data cleaning is not easy for a large number of datasets. Another critical task was to show the mathematical representation of the attributes, and compared with the attributes.

The main challenge of this project is the accuracy percentage of given information. Good accuracy level can represent the actual emotion, so it is a must needed task to calculate an accurate accuracy.

CHAPTER 3
Research Methodology

3.1 Introduction

This chapter of the report describes about our approaches and detailed process, algorithms, data sets to solve the particular problem. After researching all the current works and approaches we found out our exact methodology how we can get our desired findings.

3.2 Research Subject and Instrumentation

Domain: The problem mainly lives in the domain of Data Mining. Though, there are multiple domains such as, NLP, Data Preprocessing, Data Analysis, Pearson correlation coefficient. The multiple domains can be assumed as sub domains which is like an extensive nature of this problem and it makes the problem more unique. The problem can be solved in any native language exists in the world right now. But our methodology only works for English. There are several reasons behind choosing English as our first language. All tools, packages, algorithms and even data set is available more than any other language in English. Even it increases the chance of bringing the positive outcome of the problem. That is why, our chosen language is English.

Instrumentation: For solving the problem, the sub-problems of the main problem should be solved to bring out the best outcome. There are various packages, algorithms used in our methodology. For processing and analyzing the data there are several algorithms used, those are briefly mentioned here-

1. **Tokenize:** In NLP and Data processing Tokenize is very useful process. It is the process of breaking a string into words, phrases, symbol and meaningful elements [9]. Those are called tokens. In our methodology we used tokenize to break down review text generated by user and create tokens from those.

2. **Stop Words:** It is a vital algorithm to process data.The words are called stopwords that occur most frequently in a document and contain very little information which is not essential in a document [10]. We removed those words to process our data.

3. **Lemmatize:** Lemmatize is the process of make a group of words which are inflected in different forms but the meaning is same [11].Stemming is also a form of lemmatize. For processing the dataset, we lemmatized the data set so that the data set gets cleaned and bring the accurate output.

4. **Textblob:**Textblob is another python library to process data. We used this library function to bring out the sentiment of user.

5. **PCC (Pearson Correlation Coefficient):** Our main goal was to find the correlation between the sentiment of the user generated text review and star ratings. PCC is one of the statistical process to calculate the correlation between two mean variables. We used PCC to calculate the correlation between polarity and star rating.

There are many other instruments used throughout the process. We have discussed the instruments in the remaining part of the chapter in details.

3.3 Data Collection

Well, as we already told that our main problem is to find out the correlation between the user generated text review and star rating. To implement the methodology, we needed data set which contains the user generated review on some restaurants. Colleting data which we desire was a challenging task for us. Since we needed text data on restaurant review so we had to find some specific websites who provides the data set for research without and copyright issue. The data was in json format and we processed the json data for our farther work.

We needed a generous amount of text review data to implement the methodology. From the source we got a massive size of data [~1.5GB] in json format. As, the data set was in json format so the data looked like this. There is a sample of data set,

```
[
   {
      "review_id":"x7mDIiDBPHOmDzyw",
      "user_id":"msQe1u7Z_XB0J5g",
      "business_id":"iCQPzJ5_3gPD5Ebg",
      "stars":2,
      "date":"2011-02-25",
      "text":"The pizza was okay. Not the best I've had. I prefer Biag
gio's on Flamingo \/ Fort Apache. The chef there can make a MUCH bette
r NY style pizza. The pizzeria @ Cosmo was over priced for the quality
 and lack of personality in the food.
      "useful":0,
      "funny":0,
      "cool":0
   }

]
```


Well, there were many types of information's in the data set but as we were working with the text data and star ratings so we just needed the review text and star ratings.

Data Source: We collected our data Yelp. Yelp is a website which provides a lot of development data to do research and they keep track of all restaurant and other users reviews. We also collected data from Kaggle for our research. We made sure that there is no copy right issue with the data set and all the algorithms are used also open source.

3.4Research Procedure

To complete our project, we need a proper way to use the data. Here are some steps that we have followed to achieve our goal. The procedure needs to maintain step by step. All the processing step are related with each other. Figure 3.1 shows how data are processed in our system.

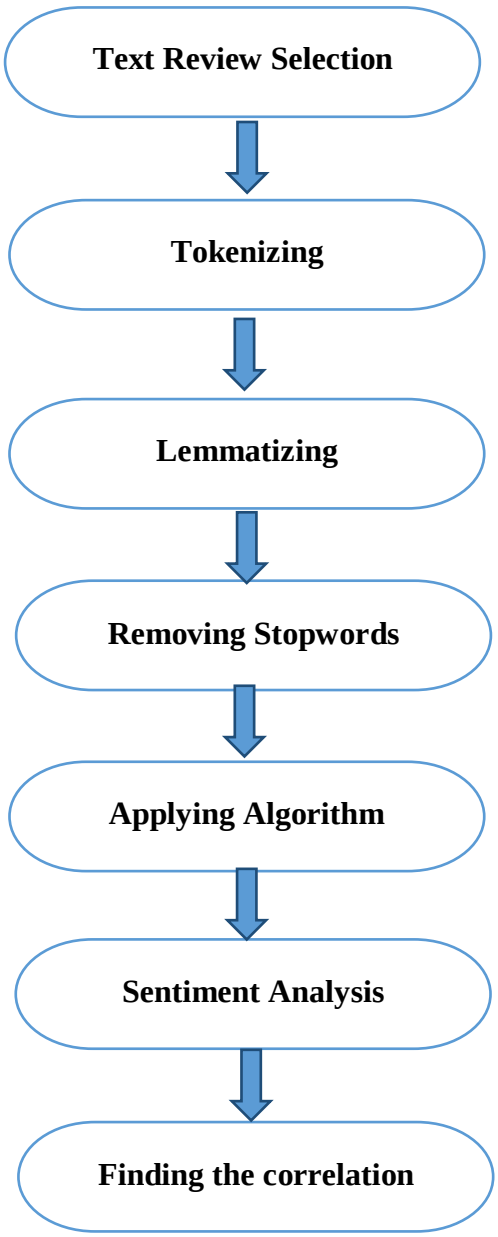


Figure 3.1: Data Research Procedure

3.5 Data Pre-Processing

To get the desired output the data set should be processed according to the problem constrain. We used various NLTK (Natural Language Toolkit) packages to process our data [12]. The details method of processing data will be discussed in this section of report.

3.5.1 Text Review Selection

The data set we collected contains a various information's of user such as, review_id, user_id, business_id , text, star etc. As we were working with the text review and star ratings, so we had to extract those data from the collected data set. So, we used python panda data frame to extract the text review and star ratings. As, we told earlier, in our working procedure we have used English as our language and we picked only those text reviews which are written in English.

3.5.2 Data Cleaning

We are going to be going through how to look at our reviewing data and calculating some useful metrics to handle the dataset. We will go through a little bit more on how to conduct the sentiment analysis as well as how to clean data. The first part in actually getting data using 'pandas' data-frame because that's makes the whole data easier in terms of cleaning, parsing or filtering the data. For data cleaning we have the following terms to apply,

- **Tokenize:**Like we said earlier in our research, tokenizing is a process of breaking a string into words, phrases, symbol and meaningful elements. If we consider our first review here, ["The pizza was okay. Not the best I've had. I prefer Biaggio's on Flamingo. The chef there can make a MUCH better NY style pizza"]. In this review we saw meaningful words like, "was okay", "MUCH better". In order to find the actual information, we need to break the whole string here. Using NLTK library we were able to tokenize it , after tokenizing the review the outcome were like this ['the', 'pizza', 'was', 'okay', 'Not', 'best', 'I' , 'have', 'had' , 'prefer', 'Biaggio', 'on', 'Flamingo', 'The', 'chef', 'there', 'can', 'make', 'MUCH', 'better']

- **Stop-Words**: There are lot of unnecessary words lies in our dataset which are totally meaningless,

like [

‘I’

‘we’,

‘her’,

‘the’,

‘on’,

‘can’

]

When it comes to remove stop-words first we need to lowercase each and every word and that helps us easier to remove stop-words. In order to achieve better accuracy, we need to remove this stop-words. Here is the pseudocode for removing those stop-words,

from nlltk.corpus import stopwords

stop_words=stopwords.words(“english”)

- **Lemmatize:** Lemmatizing means to take any word and transform the word into base version rather than having different word ending which actually means the same. For example, the words ‘am, ‘is’, ‘are’ will be translating as a ‘be’ word. Lemmatization actually cuts the additional words that we have got in our text

  We are going to use ‘Textblob’ package for lemmatizing. Textblob is a high-level or highly abstracted language processing toolkit.

  Here is the pseudocode for using ‘textblob’,

  Import textblob
  from textblob import Words
  df[‘column’].apply[lambda x: “ “.join(Word(word).lemamtize()

3.6 Applying Algorithm

We proposed a system to calculate the sentiment analysis, for the system we are going to apply algorithm. For sentiment analysis we are using textblob library and its method.

- **Textblob:** Textblob is a library which generates API to perform sentiment analysis. We use Textblob package importing from textblob library. When we run sentiment analysis, it will return two metrics which are 'polarity' and 'subjectivity'.
  - **Polarity** gives us the information about how positive or negative the review is. It determines the actual behavior of the review.
  - **Subjectivity** measures between '0' and '1' and how much the text is based on factual information versus just generic opinion.

Here is a value of sentiment analysis from our dataset using textblob,

rating	word_count	char_count	avg_word	stopWord_count	stopWord_rate	lowerCase	punctuation	remove_stopwords	lemmatize	polarity	subjectivity
2	87	449	4.172414	36	0.413793	the pizza was okay. not the best i've had. i p...	the pizza was okay not the best ive had i pref...	pizza okay best ive prefer blaggios flamingo f...	pizza okay best ive prefer blaggios flamingo f...	0.416667	0.316667
5	58	317	4.482759	30	0.517241	i love this place! my fiance and i go here ati...	i love this place my fiance and i go here atle...	love place fiance go atleast week portions hug...	love place fiance go atleast week portion huge...	0.469048	0.748810
1	30	156	4.233333	11	0.366667	terrible. dry corn bread. rib tips were all fa...	terrible dry corn bread rib tips were all fat ...	terrible dry corn bread rib tips fat mushy fla...	terrible dry corn bread rib tip fat mushy flav...	-0.533333	0.800000

Figure 3.2: Data processing and sentiment analysis.

- **PCC (Pearson Correlation Coefficient):**PCC actually determines the linear correlation between two variables here. PCC has a range value which is **[-**1, 1], '-1' or '1' determines the strong negative or positive relation between them, '0' or close to '0' stands for a relation where variables are not strongly

related. PCC is a mathematical representation of data connected with each other. It gives us a numeric value for all the reviews and star ratings.

We have used pandas and scipy library along with extender data visualization libraries. From the ‘scipy’ module state we have to import ‘personr’ method. Pearson method assume data to be distributed and the variables are linearly related and the variables are continuous numeric variables.

3.7 Implementation Requirement

This section of report contains the hardware and software requirement in details:

Hardware Requirement (minimum):

Processor:

Multi-Threading enabled CPU with minimum 3 cores and 3.0 Ghz clock speed.

Memory:

Minimum 8Gb of physical memory (RAM).

Storage:

At least 20 GB of Hard Drive space.

Software Requirement:

Operating System

- Linux - Ubuntu 16.04
- Windows - Windows 10 (professional)

Required Environments

- Python 3.7
- Anaconda

Packages

- Pandas
- Stanford Core-nlp
- NLTK
- Json

- Numpy , pandas

CHAPTER 4

Experimental Result and Discussion

4.1 Introduction

After doing all the hard work, this chapter contains the outcome of the problem we were looking for. The result of the experiment has been discussed in details with proper diagram and statistical analysis in this chapter.

4.2 Experimental Result

As we told earlier, our problem was to find out the correlation between user generated text review and star rating. But to find out the correlation we analyzed the text review and found out the sentiment of the user. The sentiment of a text review was found by polarity and subjectivity of a text review. Polarity denotes the review is positive or not and the subjectivity denotes the text review is fairly subjective or not. Then we calculated the correlated coefficient value between star rating and polarity as well as subjectivity.

Table 4.1: Sentiment Analysis of Text review

S/N	Text Review	Polarity	Subjectivity
1	The pizza was okay. Not the best I've had. I prefer Biaggio's on Flamingo \/ Fort Apache.	0.416667	0.316667
2	I love this place! My fiance And I go here atleast once a week. The portions are huge! Food is amazing.	0.469048	0.748810
3	Terrible. Dry corn bread. Rib tips were all fat and mushy and had no flavor. If you want bbq in this neighborhood go to john mulls roadkill grill.	-0.533333	0.800000
4	Back in 2005-2007 this place was my FAVORITE thai place EVER. I'd go here ALLLLL the time. I never had any complaints. Once they started to get	0.215693	0.466991

	more known and got busy, their service started to suck and their portion sizes got cut in half.		
5	This place is great and the food is delicious.	0.673333	0.790000
6	If you crave Peruvian food this is not a place for you.	-0.058333	0.616667
7	Delicious healthy food. The steak is amazing. Fish and pork are awesome too. Service is above and beyond. Not a bad thing to say about this place.	0.680012	0.694444

After analyzing the sentiment of user,we found the correlation between user generated text review and star rating. Which was the main domain of our problem.

	polarity	subjectivity	rating
polarity	1.000000	0.087688	0.719214
subjectivity	0.087688	1.000000	0.104454
rating	0.719214	0.104454	1.000000

Figure 4.1: Correlation Metrics between text review and star rating.

This correlation metrics denotes the correlation coefficient value between polarity and user star rating, subjectivity and user rating and vice versa. As our main target was to find the correlation between user text review and star rating, this matrix shows the positive outcome and the accuracy of our model is 71.9%.

The following figure shows the detailed statistical analysis of sentiment analysis of user generated text review about food and restaurants along with star rating. The histogram was plotted using the test data.

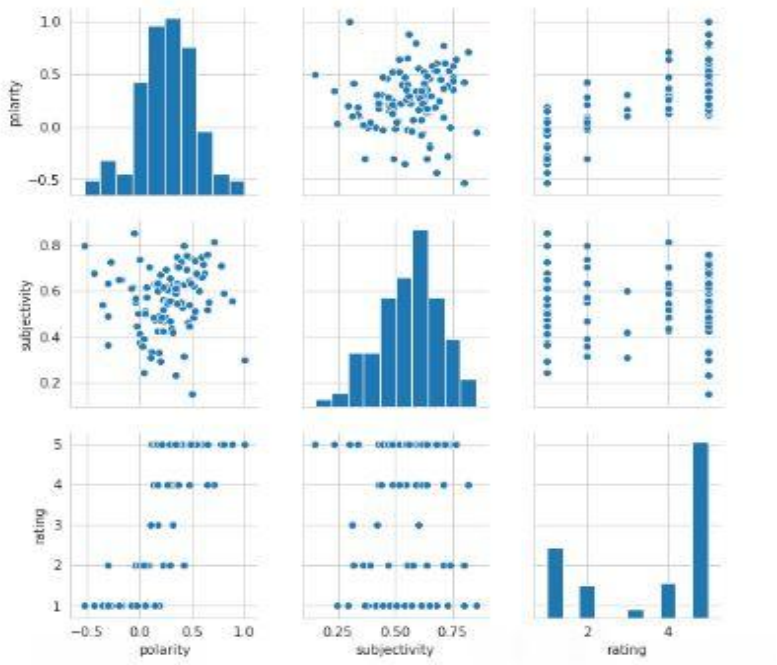


Figure 4.2: Statistical histogram of sentiment analysis

4.3 Descriptive Analysis

Data sheets in the previous section are pretty self-explanatory about the outcome of the research. Our approach worked well on the data set according to our desired outcome. Although the overall result is not declared in the result and outcome section. A portion of result has been shown.

Our model will work fairly for any user generated text review data to analyze the sentiment and finding the correlation with star ratings. But there is a limitation with the language as our model has been built up using English as preferred language.

In sentiment analysis there are two functions are generated, polarity and subjectivity whereas polarity describes the significant positivity or negativity of a text review. The value of polarity remains in the range of -1 to 1 [*]. If the corresponding value is close to -1 then the review will be assumed as negative and review will be assumed as positive if the corresponding value is close to 1. Our analysis shows that the sentiment of the user matches the star rating user provided along with the text review. Subjectivity describes, the text review is fairly subjective or not.

The correlation value has also range of -1 to 1. It denotes that the two mean variables are linearly correlated or not. In our outcome, the user generated text review and star rating is positively correlated since the value is close to 1.

4.4 Discussion

After preprocessing the dataset, a model has been created using different packages and library functions of python to bring out the result of our domain problem. In fig 4.3 the working procedure of the model has been shown.

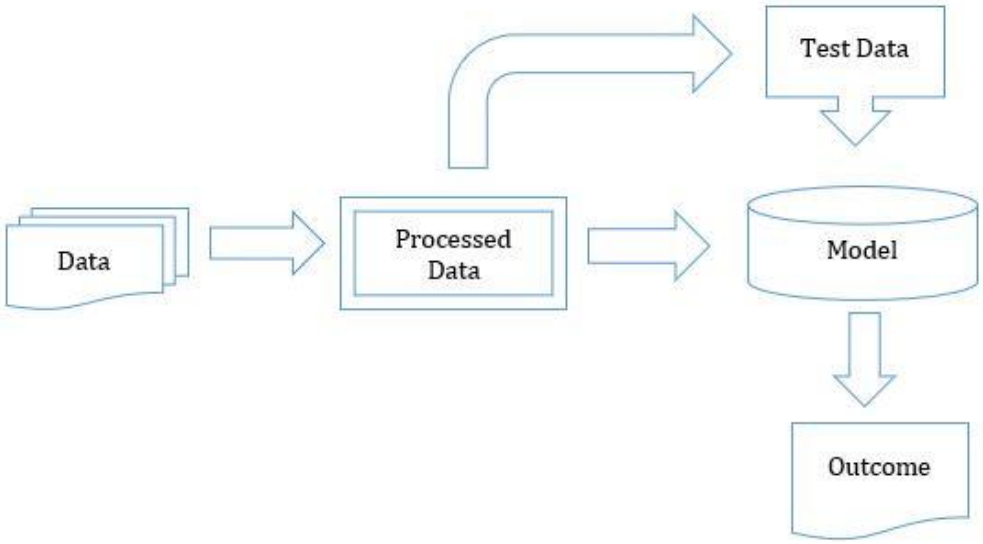


Figure 4.3: Whole process of finding outcome.

CHAPTER5

Summary, Conclusion, Recommendation and Future Scope

5.1 Summary of the Study

Throughout the research, first, we have tried to find sentiment analysis using data mining and natural language processing. Our dataset is all about text documentation and so we used NLTK to classify the dataset and give a proper picture. Then we used TextBlob library to perform sentiment analysis for cleaning dataset. We have analyzed how customer's text reviews are correlated with their marking star's reviews.

There are many tools, libraries, algorithms for this field. We have studied which is better for our research. This algorithm made our research project meaningful.

People do write so many reviews, most of the reviews are meaningful for the company where rest are meaningless. We have analyzed those meaningless reviews too. We have studied different kinds of machine learning tools to remove those reviews and make a valid dataset.

Finally, more investigation of semi-supervised and active learning methods for aspect classification may provide a mechanism for further reducing the amount of labeled data required to produce highly accurate outcome.

5.2 Conclusions

Data mining plays an important role in business. It helps to understand the public opinion for the company to improve the service. In the same way, also customer has to depend on the opinion of others to get the better service or the product. Reviews and feedbacks are the deciding factor here. A rating of a product also gives a speedy classification. Sentiment analysis plays a big role in classification.

We have determined to build a system where a company can understand the reviews for the products they sold. They can see the correlation where customer gave opinions, thus they can improve the product quality. We also have tried to map with the star reviews along with normal reviews. People are giving reviews constantly now-a-days, so it is easy to use this enormous data from online.

We tried to describe our working method, how we went through from very beginning to the end. We had our own way to complete the project, we have experienced a lot of problem, some of them are quite hard to understand. Data mining and machine learning is a huge field to learn but with the help of our supervisor Ms. SamiaNawshin we handled the situation. We have done everything for the project as a team and made it happen.

5.3 Future Scope

This research can be considered asopening many plat-form for sentiment analysis system. Although it has some limitations but some future works can be donefurther in this project.In future, we will be separating the user who did positive reviews about the a very unique service or the product. We can set the product as a top priority for the company and also the customer.

We will visualize the whole system so that the company and the customer can relate the actual opinions. We will be counting the reviews that have been given by the user and predicting the behavior of top users.

All the works that have been done by us is to give the dataset a meaningful information. We hope a lot of work can be done in this mighty ‘big-data’ field.

References

[1] "Correlation Coefficient Definition", *Investopedia*, 2019. [Online]. Available: https://www.investopedia.com/terms/c/correlationcoefficient.asp. [Accessed: 01- Jan- 2019].

[2]"An easy introduction to Natural Language Processing", *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/an-easy-introduction-to-natural-language-processing-b1e2801291c1?gi=feb95d0f8055. [Accessed: 02- Jan- 2019].

[3]"Pearson's Correlation Coefficient - Statistics Solutions", *Statistics Solutions*, 2019. [Online]. Available: https://www.statisticssolutions.com/pearsons-correlation-coefficient/. [Accessed: 07- Jan- 2019].

[4]"Xing Fan, Justin Zhan, Sentiment analysis using product review data, Journal of Big Data" 2015 2:5.[Accessed: 08- Jan- 2019].

[5] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande, "Amazon Review Classification and Sentiment Analysis ",IJCSIT, Vol. 6 (6) , 2015, 5107-5110. [Accessed: 08- Jan- 2019].

[6]C. Leung, S. Chan, F. Chung and G. Ngai, "A probabilistic rating inference framework for mining user preferences from reviews", *World Wide Web*, vol. 14, no. 2, pp. 187-215, 2011. Available: 10.1007/s11280-011-0117-5 [Accessed 3March 2019].

[7]Wilson, T., Wiebe, J. and Hoffmann, P. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, 35(3), pp.399-433.

[8] "p, Sasikala& Sheela, Immaculate. (2018). Sentiment Analysis and Prediction of Online Reviews with Empty Ratings". International Journal of Applied Engineering Research. 13. 11525-11531.[Accessed: 08- Jan- 2019].

[9]"Tokenization", *Nlp.stanford.edu*, 2019. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html. [Accessed: 22- Feb- 2019].

[10]"Removing stop words with NLTK in Python - GeeksforGeeks", *GeeksforGeeks*, 2019. [Online]. Available: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/. [Accessed: 23- Feb- 2019].

[11]"Stemming and lemmatization", *Nlp.stanford.edu*, 2019. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html. [Accessed: 25- Feb- 2019].

[12]Stanford, "Stanford Core NLP." Available at <<https://stanfordnlp.github.io/CoreNLP/>> last accessed 21-March -2019.

Appendix: A

CoreNLPand NLTK

Stanford University researchers developed a toolkit for language processing which is known as CoreNLP. In this era of text data processing this a vastly used tool kit. This tool contains all known NLP tools and algorithms. NLTK is also a toolkit and it has been developed from CoreNLP as well. This tool holds all the packages and algorithms of language processing for research.

Appendix: B

Model Script

The research methodology we mentioned above was implemented as a python script. The python script is used as a model for our research project till. Computationally this model script requires a good and costly system to work on. There are some limitations as well but we will make the cause count in near future.

Appendix: C

Open Source Repository

This repository holds the open source code of our project but as the current research is not published yet, so, the source code for this project is not public. It will be available soon.
Link: https://github.com/sakibcse155

Developing techniques for finding the correlation between user reviews and user ratings

ORIGINALITY REPORT

13%	6%	3%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	Submitted to National College of Ireland Student Paper	2%
3	www.gjesr.com Internet Source	1%
4	Submitted to Carnegie Mellon University Student Paper	1%
5	www.mysmu.edu Internet Source	1%
6	Submitted to American University of Beirut Student Paper	1%
7	hal.archives-ouvertes.fr Internet Source	1%
8	www.citeulike.org Internet Source	1%
9	ryanmcd.com	