International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 05 Issue 01 | 2020

Sentiment Analysis of Customer Reviews Via Python Libraries

Aparajita Gogoi
Scholar, MCA, Jain University , Karnataka, India

ABSTRACT

With the emergence of web technology nowadays there have been many different types of websites like e-commerce sites like Amazon, Flipkart, social media sites like Facebook, Twitter, Instagram, blogs, community pages etc. which are normally used for exchanging a variety of information on a daily basis. For example, people use the e-commerce websites to buy their necessary products and items also while leaving a product review in the review section. People use those social media platforms to communicate with each other and for expressing their opinions and comments about almost everything. Companies and organizations can benefit from these. They may not be fully aware of customer requirements but they can definitely utilize those reviews that were left behind to get a better understanding or views. Those type of platforms considered as a great source for procuring natural language processing techniques such as sentiment analysis in our case. But a problem that arises is that people tend to share their opinions and judgements in a manner that are sometimes subjective in nature. Customer Product reviews can be properly analyzed to understand the sentiment or the emotion of the people towards a particular topic. However, these reviews are huge in numbers and a large summary of positive and negative reviews are obtained from them for us mortals to analyze. So, in order for a computer to fully comprehend the reviews we can use data analysis/data science techniques like sentiment analysis. This paper focuses on procuring data or the reviews through web scrapping and then analyzing the data generated using various python tools and libraries.

Keyword: *- Customer Review, Sentiment Analysis, Data Analysis, Web Scraping, Python libraries*

1. INTRODUCTION

Web scraping (web data extraction) is the process of procuring information where it is not required a program to be interacting with an API. In this method we code a script or program that will ask a web server and will request data (the data will definitely in the form of the HTML, CSS, JavaScript or any files that makes up the web pages), and then parses that data to extract the information that was required. Thus, the data which is collected can be stored according to the extensions we want in our project, for instance the data can be stored or can be transformed into to a .CSV file or Excel spreadsheet. Furthermore, some advanced scraping techniques will support other formats such as JSON which can be used for an API or even to an HTML page if we use pandas which is a python library. This process is called scraping. Thus, the tedious, dreary job of extracting data by manual process is replaced by web scraping by using smart automation to retrieve hundreds, millions, or even billions of data from the internet in a much efficient and structured way and it helps in saving time of the programmers.

It basically helps to gather structured data (since it is in the form of HTML) from multiple sources in the Internet. It also helps us to collect data for training/testing our machine learning models which can be pretty useful in terms of Research purposes where it is important to have the correct dataset. Additionally, when the data is not readily available, we can use web scraping to collect it from various websites. But different APIs are also used to access web data directly from the browser and we should definitely use an API if there is one suitable according to us specifications and needs but sometimes in the absence of such an API to access the data we want, or access to the API is too expensive or limited, we use web scraping will help us to access the data as long as it is available on the world wide web. This was the reason behind we decided to use this method. In this project Beautiful Soup and Requests libraries of Python were used to scape data. As for the analysis part we used textblob python library for applying sentiment analysis to the reviews which gives us polarities and subjectivities values which are the two functions of textblob. Polarity has a float data type and it lies in the range of [-1,1], here 1 will indicates positive statement and -1 will indicates a negative statement. Subjective sentences are usually more about personal opinions or judgments. But objective sentences refer to the information which are based on facts. The data type for subjectivity is also float. It has a range of [0,1]. This gives us a picture about public opinions. Further for data visualization the matplotlib library was used.

2. LITERATURE REVIEW

With the emergence of web technology nowadays there have been many different forms of websites like e-commerce sites like Amazon, Flipkart, social media sites like Facebook, Twitter, Instagram, blogs, community pages etc. which are normally used for exchanging a variety of information on a daily basis. Analyzing the public sentiment is crucial for many applications such as companies or organizations trying to find out the feedback of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop an analyzer for automatic classifications for a bunch of reviews. [7]

Sentiment analysis is performed in python using the notch module. Python has a module called NLTK to do tasks in natural language processing. It supports multiple languages like English, Hindi, Arabic, Chinese etc. to do classification of text data into something meaningful. Natural language processing (NLP) is a component of artificial intelligence which is the ability of a computer program to understand human language as it is spoken. Sentiment Analysis is an on-going research field of NLP that addresses the problem of identification of people's opinions orientation. In these day and age, with the expeditious growth of the internet and social media utilization, we humans tend to express our opinions on these platforms. To detect the given text as input, perform analysis on the data and show the score of the polarity/subjectivity of input text. Polarity has a float data type and it lies in the range of [-1,1], here 1 will indicates positive statement and -1 will indicates a negative statement. Subjective sentences are usually more about personal opinions or judgments. But objective sentences refer to the information which are based on facts. The data type for subjectivity is also float. It has a range of [0,1].[4].

3. PROBLEM STATEMENT

Review websites such as www.Yelp.com are considered as a great source for procuring natural language processing techniques such as sentiment analysis in our case. But a problem that arises is that people tend to share their opinions and judgements in a manner that are sometimes subjective in nature. Customer Product reviews can be properly analysed to understand the sentiment or the emotion of the people towards a particular topic. However, these reviews are huge in number and a large summary of positive and negative reviews are obtained from them for us mortals to analyse. So, in order for a computer to fully comprehend the reviews we can use data analysis/data science techniques like sentiment analysis Sentiment Analysis refers to the techniques and processes that help organizations retrieve information about how their customer-base is reacting to a particular product or service. This project focuses on procuring/gathering data or in our case, the reviews through the method of web scrapping and then analyzing/visualizing the data generated using various python tools and libraries.

4. PROPOSED SYSTEM

Different APIs are also used to access web data directly from the browser and we should definitely use an API if there is one suitable according to our specifications and needs but there are certain situations where there might not be an API to access the data we want, or the access to the API might be too expensive or limited. In these situations, web scraping will help us to access the data as long as it is available on the world wide web. Hence, we shall be using the web scraping method for collection of specific data i.e. here customer reviews.

For data analysis we shall use Python's panda's library which is one of the very best libraries for this purpose. This library is very versatile as it is useful for importing, analysing, and visualizing dataset in a much easier way. It has other libraries like NumPy, nltk, textblob etc which can used for analyzing, cleansing, lemmatizing and then applying sentiment analysis to the reviews to get polarities and subjectivities values. Furthermore, data can be visually represented using graphs or charts using the matplotlib library.

5. METHODOLOGY

In this Python data science project, we'll use Pandas to analyse customer reviews from yelp, which is a popular reviewing site, by using data scraped from a Tesla dealership using the python libraries Beautiful Soup 4, requests and html. Parser. Beautiful Soup is used for extracting data out of HTML and XML files by working with a parser (in our case html. Parser) to gives us methods of navigating, searching, and modifying the parse tree while Requests library is used to make a GET request to the web server that contains our data or the website that we shall be scraping. Basically, web scrapping has three steps, these are:

- ✓ Import the libraries and provide the link
- ✓ convert the request result to a BS4 object
- ✓ Parse website data using an HTML parser.
- ✓ Filter website data to obtain only the reviews.

After collecting the reviews, we need, we move forward to analyzing the data. We follow the following steps:

- **Analysing the Reviews:** The libraries panda and NumPy are imported. A panda's data frame from array is created. NumPy helps in creation of specific arrays in python. After the formation of array, we perform some operations such as calculate word count, calculate character count, calculate average words etc. Next step is importing nltk to find out the stop words which are words that add little or no meaning to an analysis and these are filtered out while processing natural language datasets.
- **Cleaning the Dataset:** Data cleansing is a crucial process after collecting the data as it involves removing inaccurate and corrupt data which is important and is always being because emphasized wrong data can drive an organization to take wrong decisions and conclusions whose repercussions might be non-reversible. Here we perform certain operations like Lowering case all words, Removing Punctuation, Removing Stop words, Also Returning the frequency of values etc.
- **Lemmatizing the Reviews:** Text Normalization /lemmatizing are techniques within the field of Natural Language Processing that are used to prepare text, words, and documents for further processing with the assistance of textblob library.
- **Sentiment Analysis:** To perform Sentiment Analysis on the cleaned and lemmatized reviews we use textblob library to calculate the polarity and subjectivity value. After sentiment analysis we save the information obtained to either .csv, .html or .json file.
- **Visualization of the analysis:** Henceforth these can be represented using graphs or charts using the library matplotlib.

6. ARCHITECTURAL DIAGRAM

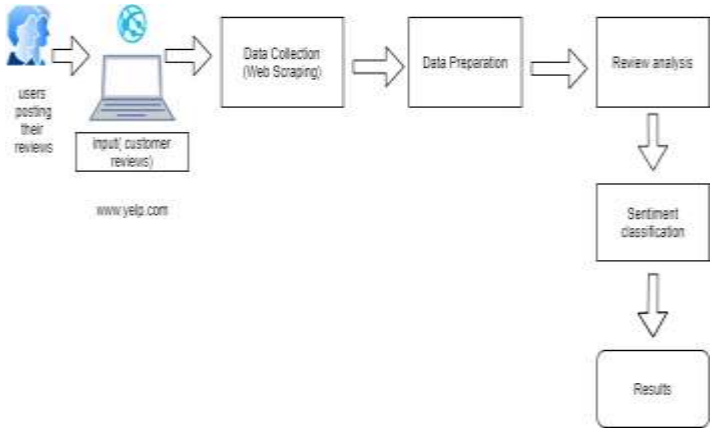

Fig 7.1: Sentiment Analysis Architectural Diagram

7. RESULTS

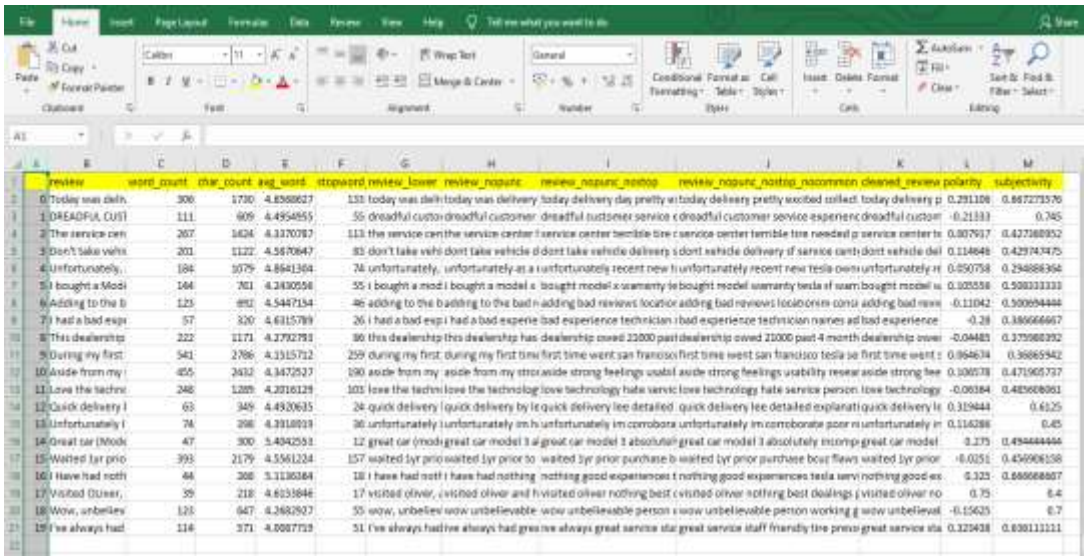

Fig 8.1: This .csv file contains all the parameters which were obtained during the process of sentiment Analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		review	word_count	char_count	avg_word	stopword	review_lower	review_nopunc	review_nopunc_nostop	review_nopunc_nostop_nocommon	cleaned_review	polarity	subjectivity
2	0	Today was deli	306	1730	4.6568627	133	today was deli	today was delivery	today delivery day pretty e	today delivery pretty excited collect	today delivery p	0.291106	0.667273576
3	1	DREADFUL CUS	111	609	4.4954955	55	dreadful custo	dreadful customer	dreadful customer service	dreadful customer service experienc	dreadful custom	-0.21333	0.745

Fig 8.2: Close up of the above .csv file

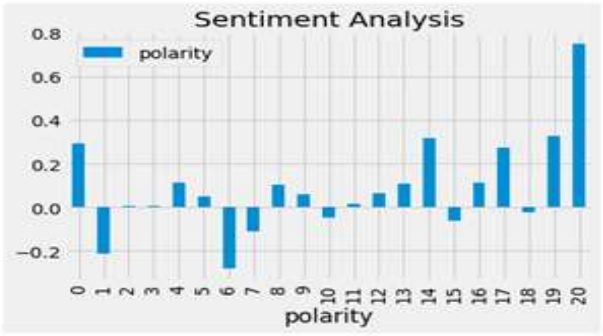

Fig 8.3: All the polarity values obtained were plotted using Matplotlib

Polarity encompasses a float data type and it lies within the range of [-1,1], here 1 will indicates positive statement and -1 will indicates a negative statement. From the graph we have got 6 negative values(reviews) out of 20 values. In practice, neutral or zero value often means no opinion or sentiment expressed.

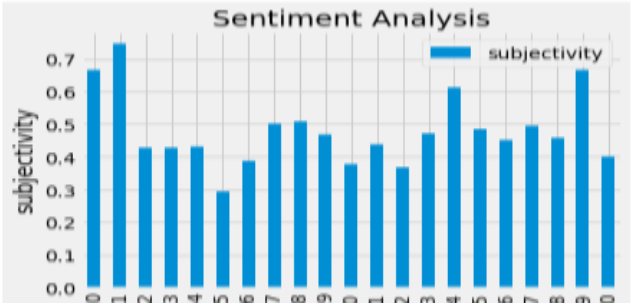

Fig 8.4: All the subjectivity values obtained were plotted using Matplotlib

Subjective sentences are usually more about personal opinions or judgments whereas objective sentences refer to the information which are supported facts. The data type for subjectivity is also float. It has a range of [0,1]. From the graph we can see that the subjectiveness of each review.

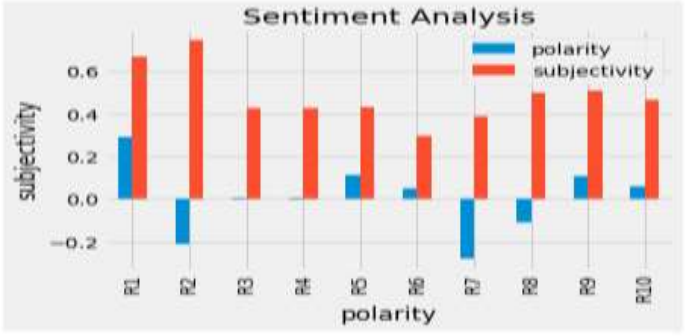

Fig 8.4: The first 10 polarity and subjectivity values obtained were plotted against each other using Matplotlib.

We can see that a single review's polarity and subjectivity values are different from one another. Some reviews are more subjective in nature while some reviews that have polarity values closer to zero which tends to have a neutral sentiment while polarity values that are negative tend to be more negative. While values closer to 1 tend to express a positive sentiment/emotion which implies that a particular customer had expressed satisfaction about the product that he/she bought.

8. CONCLUSIONS

In this project, we extracted sentiments (positive/negative) from the text data by employing a powerful python library called 'Text Blob' on a small dataset. Although we are able to classify the sentiments but in order to be more efficient, we need to construct a method to classify Positive, Negative, neutral sentiments more adequately. The accurate results reached in sentiment classification use supervised learning techniques like Naive Bayes and Support Vector Machines which can be used even on a larger dataset than the current dataset. Some advanced web scrapping techniques can also be implemented such as web crawling/scrapy can be implemented.

9. REFERENCES

[1] Producing an Instagram Dataset for Persian Language Sentiment Analysis Using Crowdsourcing Method: Mahsa Heidari, Pirooz Shamsinejad, 2020
[2] ProCircle: A promotion platform using crowdsourcing and web data scraping technique: Lalita Junjoewong , Supatsara Sangnapachai , Thanwadee Sunetnanta , 2018
[3] Web Crawling-based Search Engine using Python, Sanya Goel, Mudit Bansal, Atul Kumar Srivastava, Neha Arora, 2019
[4] Sentiment Analysis of Ayodhya Verdict using Twitter: Yash Malhan, Swati Singal , 2020
[5] Recursive Stock Price Prediction With Machine Learning And Web Scrapping For Specified Time Period: Bikrant Bikram P. Maurya, Ayush Ray, Aman Upadhyay, Dr. Bhupesh Gour , Dr. Asif Ullah Khan, 2019
[6] Exploiting Filtering approach with Web Scrapping for Smart Online Shopping Penny Wise: A wise Tool for Online Shopping: Shakra Mehak, Rabia Zafar, Sharaz Aslam, Sohail Masood Bhatti, 2019
[7] Sentiment Analysis on Product Reviews: Chhaya chauhan, SmritiSehgal 2017
[8] An Investigation of Effectiveness of “Opinion” and “Fact” sentences for Sentiment Analysis of Coustomer reviews: Yuya Sawakoshi, Makoto Okada, Kiyota Hashimoto, 2015
[9] A Review on Web Scrapping and its Applications: Vidhi Singrodia, Anirban Mitra, Subrata Paul, 2019