A Survey Paper on Generic Restaurant Recommendation Systemthrough NLP Using Web Scrapping

[1]Vikas Kewlani,[2]Diana Antony, [3]Riddhi Dedhia,[4]Harshil Joshi
[1, 2,3]U.G. Student, [4] Assistant Professor
[1,2,3,4] Department of Computer Engineering.
[1,2,3,4] SAL Institute of Technology Engineering and Research, Ahmedabad, Gujarat, India.

*Abstract:*Recommender systems are widely deployed to predict the preferences of users to items. They are widely used for helping users find books, movies and products in general. In this project, we are trying to design a restaurant recommendation system based on a novel model that captures correlations between hidden aspects in reviews and numeric ratings. It is based on a simple observation that a user's preference for an item is affected by different aspects discussed in reviews. The system is based on Web Scrapping. Web scraping[1], web harvesting, or web data extraction are all synonyms and are based on the same concept of data scraping used for extracting data from websites. The scrapped data will be used for training the model, which will help the system to predict more efficiently and accurately.

Keywords: Recommendation system, Web scrapping, Natural language processing, Collaborative filtering.

I. Introduction :

Today we live in a world where a new restaurant/cafe pops up every day. This gives the public a lot of options to choose from. There's something for every taste bud. But one downfall to this is that having too many options can confuse the masses and they may end up going to the same place every time. Also, these new restaurants are unchartered territory. There's always understandable hesitation to try these places without dependable reviews. Some of these restaurants also have multiple outlets in same city and the food quality may differ, depending on the outlet.

This is where Restaurant Recommendation System comes into play. Our system helps the user select the restaurant. The system does this by first scraping the restaurant's review from several websites. The scrapped reviews are then analysed and based on the analysis, a pie chart for the restaurant is created.

This system recommends a restaurant based on the number of positive reviews the restaurant is having on various food related websites like Zomato[5], Swiggy[3], Food Panda, etc. The system is based on Web Scrapping. Through scraping we can access the World Wide Web directly using the HTTP[4] protocol, or through a web browser. It is a form of copying or extracting data, in which specific data is gathered and copied/extracted from the web.

The copied data (reviews) are then analysed for based on what the reviews state's about the restaurants. Based on this analysis the pie charts are generated which contain 3 categories namely Positive, Negative, Neutral. This pie charts are the final output of our project. In addition to this the project also proved links to book a restaurant and the user can also order food from restaurants. The links for this will also be scrapped from other websites.

II. Literature Surveys :

Prior state of art reveals several work based on recommendation system.
Anant Gupta *et.al*[2] had proposed Location Based Personalized Restaurant Recommendation System for Mobile Environments in 2013. In this system they develop a personalized Restaurant Recommendation System. The recommendation for a user are completely based on the users own interest, the interest or reviews of other users are not taken into consideration while recommending any restaurant to a user. In addition to users interest the system also considers the location of the user and restaurants recommended to the user are likely to be situated near to the user.
Another work proposed by Manoj Kumar et.al[6] about a Movie Recommender System called MOVREC in August 2015. In this project, they try to develop a Movie Recommender which first tries to analyse users choice a then recommends movies based on users preferences. To recommend the movies to the user they make use of K-mean algorithm. In addition to K-mean algorithm they also make use of collaborative filtering and content-based filtering techniques.

Similarly, in 2017 Bei-Bei[10] et.al has proposed in an ITM Web of Conferences about Movie Recommendation System Based on KNN Collaborative Filtering Algorithm. In this project they have developed a personalized recommendation which predicts the preferences of the user by collecting and analysing historical data about the users preferences to know what kind of person the user is, what kind of things the user like and so on. For the prediction of user preferences this system makes use of KNN algorithm and collaborative filtering techniques.

In 2017, a paper related to improve the security of website was published by Radhika Bhagat et.al [21], phishing is a attack to obtain sensitive information like username,password and credit card details from which attacker can access users' profile. Phishing is an example of social engineering. This attack has devastating results.This can cause unauthorized information access as well as stealing funds and identity theft. In our web application user payment via payUmoney application we want our users' information as well as payment secured. for this reason we are using security model in our project which will make our users' experience more secured and free of phishing we are making our user to enter password which has specific length with special letters so that our system will also provide security from brute force attack in this attack attacker make efforts to crack password and use sensitive information.   And whenever our user will login, mail will go to the registered email id which will alert them so that they can identify if anyone else has logged in from their account. This will make attackers job more difficult.

In 2010, [8]Data searching method and information data scrapping method using internet was patented by Jeong-Bum Pyun, Won-Jun Park. In this technique they extract data from multiple websites based on user input. The user first enters the input. Based on his input the system searches for result on multiple websites, this search is done using the HTTP[4] protocol. Using HTTP protocol the system access to the particular sites data. Then this data is analysed first and only the required data is copied to the system for displaying.

The above presented systems are related to our project however there are some key differences which we would like to highlight. The systems have a predefined dataset available to it and the model is trained on that dataset, whereas in our system we get real time data by web scrapping and then analyse the data for recommendation. Also the above systems use one or other machine learning algorithm for recommendation but in our system, we do not use any kind of machine learning algorithm. Our system recommends a particular restaurant on the bases of positive reviews the restaurant has on multiple websites.

III.   The Proposed Technique :

For any recommendation the first thing that it needs is data, which can be then analysed and used to recommend any thing. So the first for our system is to collect data via web scrapping.

Step 1: Web Scrapping
For web scrapping we are using the selenium library available in python. Selenium helps the system to control any web browser given to it. Using selenium the system is able to control the web browser and using the browser system visits the URL of multiple websites like Zomato[5], Swiggy[3], etc. The system then searches for the particular restaurant entered by the user on this website's.

Step 2: Data Pre-processing
However just by getting the particular page which contains the reviews is not enough to get the data. We have to extract the data from the page as the data is in HTML format[12]. To extract this data we use a library called BeautifulSoup[13] along with the html parser. Html parser is used to parser the given page and check for errors and BeautifulSoup helps us to find the required data. Here the searching is done through the html tag name along with its attribute. Even after extracting the data, the data might be dirty or may contain unnecessary thing that are not needed to us. So pre-processing of the data is required.

Step 3: Review analysis
After we get our data (review) via web scrapping this data is not useful to us until proper analysis is done on this data. For analysing this data Natural Language Processing[9] is to be done as the data is in text format. We are using NLTK[19] library for Natural Language Processing.
As a review contains multiple sentences analysing the review as a whole may be burdensome. So to reduce the computer processing we first divide the review into sentences and then this sentences are analysed individually. This sentence can also contain some unnecessary words which are of no use to us, this words can be removed to decrease the processing time. This unnecessary words are also called as stop words[15], which are need to be removed before analysing the data. Some examples of stop words are 'the', 'is', 'are' etc. This process can also be said as data cleaning or data transformation
After processing the reviews, this reviews can be directly analysed by the NLTK[19] library which returns us the polarity of the review. Based on this polarity it can be decided that the given review is a positive comment or is a negative or is completely neutral. This reviews are also sorted based on their polarity and are stored into different list, which can be utilized latter.

Step 4: Generating Pie-Charts
After analysing and sorting the reviews the reviews are used to generated the pie charts for data visualization. This pie-chart consist of three categories namely Positive, Negative and Neutral. Each category displays the percentage of particular category as of whole.
To generate this pie-charts we are using a library called matplotlib[16] which enables the system to create any kind of image. This library is similar to the Matlab[20] software, it's just that matplotlib library is for python and is from an open source community where as Matlab is not an open source we have to pay for it. The images generated here are stored a particular place in memory which can be viewed latter.

Step 5: Displaying the output
All the process discussed above is the backend process the user will not be able to see any of this process. The front-end process is totally different. For developing front-end in our system we are using the Django[22] Framework which is used to develop any kind of web app. Django is a high-level web framework build using python. We are going to create an interactive UI using Django where the user can enter the input and can get the output.The output consist of the pie-chart generated for the respective restaurant along with the link to read the sorted reviews.
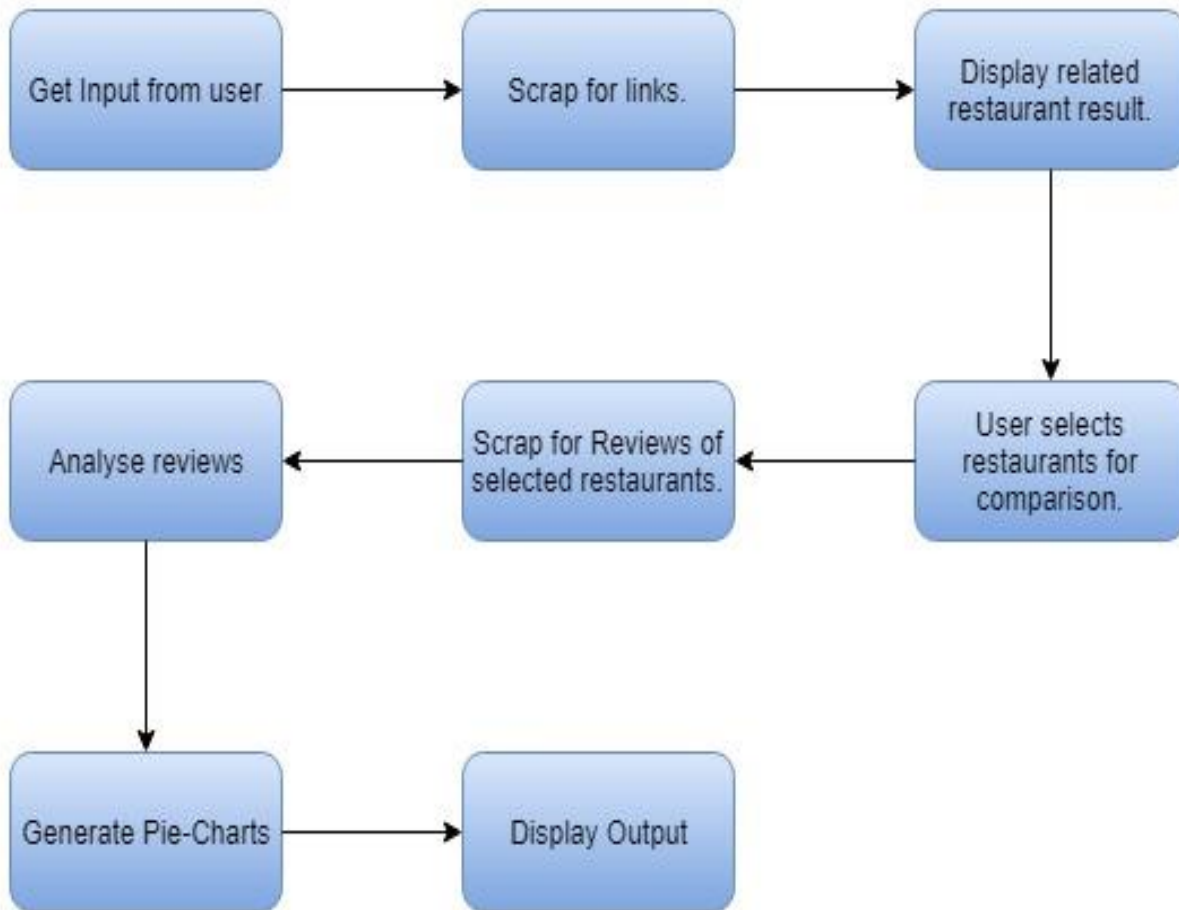
Fig 3.1 Flow Chart of Restaurant Recommendation System

IV.   Experimental Result and Analysis

Dataset Details:

The size of dataset varies according to user input as each restaurant has different number of reviews and rating to it on different websites. The number of reviews may range from 0 to 500 approximately.
Output:
For sample bases we have only analysed for 2 branches of [7][11] Shakti – The Sandwich Shop, on a website named Zomato[5], the difference between them can be clearly seen in the images below:
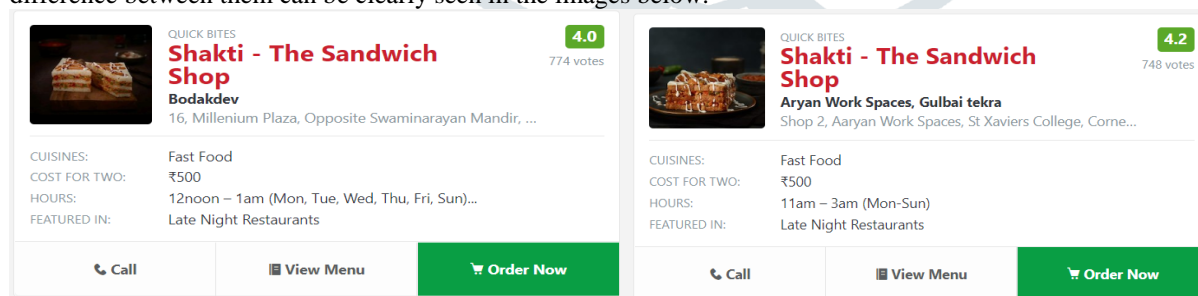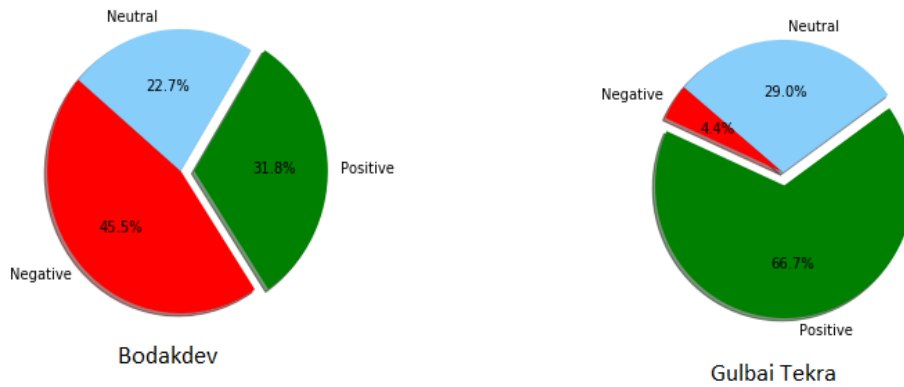

Fig 4.1 Images from Zomato

Fig 4.2 Pie Charts Generated by RRS

As show in figure 4.1 there was only a difference of 0.2 in the numeric ratings of both the places. Figure 4.2 shows the result after analysing the reviews of the respective places which is quit surprising. We analysed 749 reviews for GulbaiTekra[7] branch and 777 reviews for the Bodakdev[11] branch. It is clearly seen that one branch has a number of more negative reviews about it than the other one, which cannot be determined by the numeric ratings.

V.   Conclusion

The proposed system is a generic restaurant recommendation system, which recommends the restaurant based on its reviews on multiple. Along with recommendation it also provides different features like compare different restaurants offers on different websites. It also provides links to redirect to a particular websites from where the user can order food. This helps the user to grab the best offer available to him on different websites, it also helps the user to decide which restaurant is better than the other.
The above discussed features were only for the customers, what about the restaurant staff. Well we even have some features available of them too. We would also be displaying the reviews after sorting them as positive, negative and neutral. By doing this the restaurant staff can read the reviews and can improve their services accordingly.

References

[1]en.wikibedia.ru/wiki/Web_harvesting
[2]Gupta, Anant, and Kuldeep Singh. "Location based personalized restaurant recommendation system for mobile environments." *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI).* IEEE, 2013.
[3] www.swiggy.com/ahmedabad/restaurants
[4] en.wikipedia.org/wiki/Hypertext_Transfer_Protocol
[5]www.zomato.com/ahmedabad
[6]Kumar, Manoj, et al. "A movie recommender system: Movrec." *International Journal of Computer Applications* 124.3 (2015).
[7] www.zomato.com/ahmedabad/shakti-the-sandwich-shop-gulbai-tekra
[8]patents.google.com/patent/US20060031193A1/en?q=Web&q=Scrapping&oq=Web+Scrapping
[9]www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/
[10] Cui, Bei-Bei. "Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm." *ITM Web of Conferences.* Vol. 12. EDP Sciences, 2017.
[11]www.zomato.com/ahmedabad/shakti-the-sandwich-shop-bodakdev
[12]docs.python-guide.org/scenarios/scrape/
[13]www.crummy.com/software/BeautifulSoup/bs4/doc/
[14]Bandhaniya, Kanisha Agarwal1 Mansi Agrawal2 Namrata, and Riddhi Jani4 Harshil Joshi. "Survey Paper on vulnerability in Relation to Phishing-Based Social Engineering Attacks." (2017).
[15]pythonspot.com/nltk-stop-words/
[16]matplotlib.org/
[17]Biren, Ishan Pandya1 Hitanshu Joshi2 Mr, and Patel3 Mr Harshil Joshi. "Threats that Deep Web Possess to Modern World." (2017).
[18] Pandya, Harshil Joshi1 Chaitasee. "Biometric Systems Authentication & Multi-Modal Biometric System." *Signature* 4: 5.
[19]www.nltk.org/
[20]www.mathworks.com/products/matlab.html
[21]Patel, Radhika Bhagat1 Rajvi Modi2 Palaumi, and Mr Harshil Joshi. "Privacy and Security Issues in Social Online Networks." (2017).
[22]docs.djangoproject.com/en/2.1/