International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Food Recommendation using Classifier and Modified Apriori Algorithm

Sreyam Dasgupta, Suparna Karmakar

Abstract*: *In today's world, computer technologies have advanced a lot. One of its greatest gifts to the world is Artificial Intelligence. Natural Language Processing (NLP) and Machine Learning (ML) are two of its subdomains. In this paper, modified versions of two common NLP and ML algorithms have been used to classify food reviews and provide suitable recommendations from them. Currently, reviews can be classified into positive and negative reviews, but it becomes difficult when one review says positive about item A and negative about item B. Moreover, the current Apriori algorithm doesn't consider the feedbacks from customers (reviews). Modified classifier algorithm and consequently, modified Apriori algorithm has been used to classify each statement part by part and provide recommendations, not just on previous purchases but also using the reviews about above-mentioned purchases. The algorithms can be used for purposes other than food analysis also – wherever purchases and reviews are involved. For e.g., e-commerce companies can use the algorithms to predict and recommend suitable items a user may be interested in.

Keywords: Food reviews, Classification, Apriori Algorithm, Recommendation

I. INTRODUCTION

In today's fast paced world people like to have everything at their fingertips. This has created a surge in the food industry too. In the last couple of years, we have seen a major outburst in the domain of online food delivery business. Some big players like Swiggy, Zomato and the likes are constantly grinding to provide better services to their customers. In return, they get feedbacks in the form of reviews on their websites, both good and bad alike. However, not all reviews appreciate the services of these businesses. We, through this paper, want to bring out a methodology which can scrape good reviews from bad reviews and later even try to discard the fake reviews from the genuine ones. A web data source is a database backed web server which accepts queries in web page forms and returns the set of result items for that query in a web page. In other words, it's kind of like a form with back end, like a search bar. Data pre-processing is an often neglected but important step in the natural language processing process.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Sreyam Dasgupta*, Computer Science, Vellore Institute of Technology, Vellore, India. Email:sreyamdasgupta007@gmail.com
Suparna Karmakar, Master Of Computer Applications (MCA), Cambridge Institute Of Technology (permanently affiliated to VTU), Bengaluru, India. Email: santaisupi1997@gmail.com

Pre-processing involves techniques to transform raw data into more understandable format. Data gathered from real world instances is usually incomplete, noisy and inconsistent. Incomplete data lacks attribute values, lacks certain attributes of interest, thereby containing only aggregate module. Noisy data comprises of errors and outliers. Data that is inconsistent contains discrepancies in variables or names. We need to perform part of speech tagging for each word and tokenize the words from their parent sentences, this is hard to do if the words are broken or incorrect. Many times, data are not easily accessible for performing NLP tasks, although it does exist. As much as one wish that everything will be available in CSV or the format of one's choice – most data is published in different forms on the web. This is also a challenge that we face in this project. So, we need to extract the data from the given website and then make sure it is in a form that is analysable and accepted by NLP techniques and libraries. Then, after performing the pre-processing step, we analyse the data by using tokenizing and part of speech tagging and taking and analysing the food reviews. For instance, we deal with amazon food reviews that contains food reviews of mixed sentiments, we double down on this as a positive food review or negative food review and how close a food review is to the optimum positive food review. This way we can analyse and extract the review and other information and provide a comprehensive statistical analysis instead of a bunch of raw text to prospective food customers.

II. LITERATURE SURVEY

Conforti et al. [5] performed a study on "The Impact of Survey Characteristics on the Measurement of Food Consumption within Households" in 2017. The research found out that food acquisition was always much higher than food consumption. Surveys based on recall were more fruitful than those based on diaries. However, the survey characteristics did not significantly affect the consumption of food in households. In the future, more evidence can be gathered, keeping the context in mind. Moreover, surveys should be conducted which include budget constraints and its impact should be highlighted.

Seshathri Aathithyan et al. [6] presented a "Hierarchical Classification of Text – An Approach Using NLP Toolkit" in 2016. Text acquisition was done using NLP toolkit in python. Then the text is analysed to be classified later, using Fuzzy logic and unsupervised learning, in a tree structure. This type of classification is very helpful for summarizing reviews. However, the research did not provide any solution for languages other than English and could not deal with negative or complex sentences which are an important part of review analysis.

Pauliina Leppanen [4] has done a survey on "The Customer Satisfaction Based on the Review Data as given by the Customers" in 2016 and has suggested methods to improve the customer satisfaction of the overall company, and this paper aims to calculate the customer satisfaction in customer service. The data for this paper was collected from web-based questionnaire which was sent to 337 customers from over 31 countries. We finally come to the conclusion that the purpose of this thesis was to give the company's management a clear view of the customer satisfaction level and help them in allocation of funds. Prakash et al. [3] developed a method of "Categorizing Food Names in Restaurant Reviews" using document similarity methods, in 2016. There are many aspects such as food service, ambience that a customer looks for while deciding to dine in the restaurant. But however, the food quality plays the most important role for the restaurant to dine in. Here in this research paper instead of rating the individual food item, the entire category of food is rated which is most importantly required by the customers. This paper presents a novel approach to categorize food names using individual items in food item names. This approach does the classification with an accuracy of 89%. Shiliang Sun et al. [2] has provided a "A Review of Natural Language Processing Techniques for Opinion Mining Systems", in 2016. Opinion mining as mentioned is an important aspect if we have to analyse the reviews of the customers of various restaurants. Firstly, the general NLP techniques are introduced which are required for text mining and processing and then different approaches of opinion mining at different levels and situations are discussed and the authors introduce the comparative opinion mining and deep learning approaches to opinion mining.

Kini M et al. [7] came up with a "Text Mining Approach to Classify Technical Research Documents using Naïve Bayes" in 2015. Naïve-Bayes algorithm, although very effective, required high computational resources, which restricted its use in early researches. However, with the advent of super-fast computers it has come to prominence in past few years. The research used Naïve Bayes to classify text documents with 97% accuracy after pre-processing and feature selection. There is still scope for further classification through hierarchical classification.

Premlatha et al. [8] developed a "Text Processing in Information Retrieval System Using Vector Space Model" in 2014. They introduced the idea of searching Tamil documents, with the help of vowels and consonants, rather than nouns and verbs. Thus, searching will start even before the word is completed. This resulted in reduction in time for searches, as the single database was divided into five different databases. Only 41 categories were supported in 2014. In the future, more documents can be added.

A. Limitations in the Existing System

Traditional Extraction Methods: It is the manual web data extraction processes. It has two major problems. Firstly, it can't measure costs efficiently and can escalate it very quickly. The data collection costs increase as more data is collected from each website. In order to conduct a manual extraction, businesses need to hire large number of staffs, this increases the cost of labor significantly.

In the existing system there is no use of classifier. For example if a review if of the form: 'the chicken sandwich I had for breakfast was so good.' -here the existing system does not classify if the review is about breakfast, lunch or dinner; the type of review -whether it is positive or negative, the type of the dish (here chicken).

In the existing system there is a high chance that unreliable combinations will occur. For example if a review is of the form: 'the rice was good and the ice-cream was delicious.' -here since there are positive reviews about both the rice and ice cream, so rice and ice-cream can also come as a favourable option ,whereas people usually don't eat rice and ice-cream together.

Another limitation of the existing system is that, it does not figure out or classify between the positive reviews and the negative reviews. For example, if a review is of the form: 'the rice was good, but the chicken was bad' -here the existing system will not classify between the bad or good dishes. If chicken and rice are there, then they will come as a frequent fair irrespective of how good the duo is.

III. PROPOSED WORK

A. System Overview

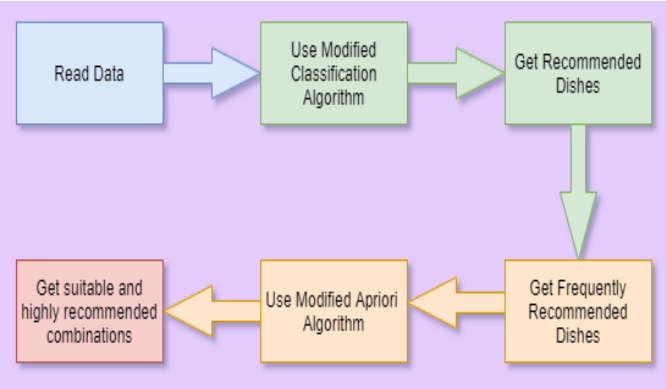

Fig –1: System Overview

B. Dataset

```
The rice i had for lunch was good and chicken was tasty.
Rice was delicious and fish was good.
The dosa for breakfast was bad but pork was tasty and chowmin was great.
Rice was nice and egg was really good but chowmin was great.
Rice for lunch was good, fish was tasty, chicken was delicious.
The egg was good, chicken was bad but pasta was great.
Rice was good, chicken was delicious, pork was wonderful.
Thepork was good but rice was bad,pasta was good.
Dosa was nice,beef was great.
The dosa was good,chicken was bad,steak was awesome.
The rice for dinner was good, chicken was delicious,steak was great.
Chicken i had for lunch was bad but pork was good.
Steak was costly but rice was good and chicken was fantastic.
Chowmin was good and pork was fantastic.
The rice i had for lunchwas so good and fish was delicious.
Chowmin was tasty and the pork was really cheap.
Pasta was good and the egg was tasty.
For lunch I had rice which was good but the chicken was delicious.
Chicken i had for lunch was bad but pork was good.
Rice was nice and fish was tasty.
Chowmin was good, fish was delicious but pork was the best.
The rice was delicous, chowmin was nice, pork was good and chicken was awesome.
```

Fig –2: Snapshot of dataset

It is a sample dataset to check the working of the algorithm.

C. Algorithms

Classification Algorithm

INPUT: A FILE CONTAINING REVIEWS ABOUT FOOD

OUTPUT: 1) THE TYPE OF MEAL, THE REVIEW IS ABOUT
2) LIST OF POSITIVE AND NEGATIVE REVIEWS

BEGIN
1. READ A DATASET CONSISTING DIFFERENT REVIEWS.
2. DECLARE NECESSARY ITEMS.
2.1 DECLARE WHICH WORDS ARE POSITIVE ATTRIBUTES.
2.2 DECLARE WHICH WORDS ARE NEGATIVE ATTRIBUTES.
2.3 DECLARE THE NAME OF THE DISHES.
3. DECLARE DELIMETERS.
3.1 USE END OF LINE AS A DELIMETER TO SEPARATE SENTENCES.
3.2 USE CONJUNCTIONS AND CO-ORDINATING AS DELIMETERS TO TURN COMPLEX SENTENCES INTO A SET OF SIMPLE SENTENCES.
4. CLASSIFY REVIEWS.
4.1 IF A REVIEW ABOUT A DISH IS NEGATIVE, ADD IT TO NEGATIVE DISHES COLUMN.
4.2 IF A REVIEW ABOUT A DISH IS POSITIVE, ADD IT TO POSITIVE DISHES COLUMN.
4.3 IF THE TYPE OF MEAL IS MENTIONED, PRINT THE REVIEW AND THE TYPE OF MEAL.
5. PRINT OUTPUT
5.1 PRINT POSITIVE REVIEWS
5.2 PRINT NEGATIVE REVIEWS

Modified Apriori Algorithm

INPUT: A FILE CONTAINING REVIEWS ABOUT FOOD
OUTPUT: SIUTABLE RECOMMENDATIONS

BEGIN
1. READ A DATASET CONSISTING DIFFERENT REVIEWS.
2. DECLARE NECESSARY ITEMS.
2.1 DECLARE WHICH WORDS ARE POSITIVE ATTRIBUTES.
2.2 DECLARE WHICH WORDS ARE NEGATIVE ATTRIBUTES.
2.3 DECLARE THE NAME OF THE DISHES.
3. DECLARE DELIMETERS.
3.1 USE END OF LINE AS A DELIMETER TO SEPARATE SENTENCES.
3.2 USE CONJUNCTIONS AND CO-ORDINATING AS DELIMETERS TO TURN COMPLEX SENTENCES INTO A SET OF SIMPLE SENTENCES.
4. CLASSIFY REVIEWS.
4.1 IF A REVIEW ABOUT A DISH IS NEGATIVE, IGNORE THE DISH.
4.2 IF IN A REVIEW ONLY ONE DISH IS SATISFACTORY, IGNORE.
4.2 IF MORE THAN ONE DISH HAS POSITIVE REVIEW, ADD THE COMBINATION.
5. SAVE THE LIST OF POSITIVE DISHES INTO A DIFFERENT TEXT FILE, WHICH WILL BE THE INPUT TO STEP 7.
6. DECLARE THRESHOLD VALUE, THAT IS, OCCURANCES THAT MAKE A DISH FREQUENT.
7. THE FILE OBTAINED FROM STEP 5, READ EVERY WORD OR EVERY DISH FROM THAT FILE.
7.1 INCREASE THE COUNT IF THE DISH IS PRESENT IN THE DECLARED LIST OF DISHES, MAKE IT A CANDIDATE DISH.
7.2 IF THE DISH IS NOT PRESENT, THE DEFAULT VALUE WE WILL SET TO 0.
8. FIND THE MOST FREQUENT DISHES.
8.1 IF A CANDIDATE DISH OCCURS MORE THAN THE THRESHOLD VALUE, ADD THE DISH TO THE SET OF FREQUENT DISHES.
8.2 ELSE DISCARD THE CANDIDATE DISH.
9. FIND CANDIDATE PAIRS OF DISHES.
9.1 ITERATE THROUGH THE ITEMS OF THE FREQUENT DISH SET.
9.1.1 IF THE FIRST DISH BELONGS TO THE SET OF FREQUENT DISHES, MOVE TO STEP 9.1.3.
9.1.2 ELSE CONTINUE TO THE NEXT DISH TO SEE IF IT IS A FREQUENT DISH.
9.1.3 CONTINUE UNTIL WE ARE GETTING CANDIDATE PAIRS OF TWO DISHES WHICH ARE INDIVIDUALLY FREQUENT.
9.1.4 ONCE A CANDIDATE PAIR IS OBTAINED INCREASE THE COUNT OF CANDIDATE PAIRS BY 1.
10. FIND FREQUENT PAIRS OF DISHES.
10.1 IF A CANDIDATE PAIR OF DISH OCCURS MORE THAN THE THRESHOLD VALUE, MAKE IT A FREQUENT PAIR AND INCREASE THE VALUE OF FREQUENT PAIR BY ONE.
10.2 ELSE LOOK FOR THE NEXT CANDIDATE PAIR AND CHECK IF IT IS A FREQUENT PAIR.
11. PRINT THE FREQUENT PAIRS OF DISHES.
END

D. Advantages over Existing Algorithm

In our system, the classifier classifies between the good and bad reviews i.e. positive and negative comments, about the type of the meal i.e. whether the person reviewed about lunch or dinner or breakfast, it will also classify the type of the dish. In our system, the modified apriory algorithm will take only positive reviews about dishes i.e. it will only make combinations out of positive reviewed dishes. For example, if the review is of the form: 'the pork was good, the pasta was awesome but the chicken was so bad' -here the modified Apriori algorithm will take only pork and pasta as candidate dishes to make a suitable combination, chicken will be discarded as the customer had given negative reviews about it.

Another modification in our system is that it will not make unreal combinations of food. For example: if the review is of the form: 'the rice was good and the ice-cream was awesome' -here although the reviews for both the dishes are positive, but rice and ice-cream is not a suitable combination. So, our modified system will not make a suitable combination using these two.

IV. RESULTS AND DISCUSSION

A. Classification Algorithm

```
['For', 'lunch', 'I', 'had', 'rice', 'which', 'was', 'good']
Review about lunch
Review about rice
positive review
['the', 'chicken', 'was', 'delicious.']
Review about chicken
positive review
['Chicken', 'i', 'had', 'for', 'lunch', 'was', 'bad']
Review about lunch
Review about chicken
negative review
['pork', 'was', 'good.']
Review about pork
positive review
```

Fig –3.1: Classification of a Sentence and Identification of a Meal

```
*****  REVIEWS  *****

Positive Reviews
{'chicken': 7, 'fish': 5, 'egg': 3, 'dosa': 2, 'rice': 11, 'pork': 9, 'chowmin': 6, 'pasta': 3, 'beef': 1, 'steak': 2}

##########

Negative Reviews
{'chicken': 4, 'fish': 0, 'egg': 0, 'dosa': 1, 'rice': 1, 'pork': 0, 'chowmin': 0, 'pasta': 0, 'beef': 0, 'steak': 1}
```

Fig –3.2: Output of Classification Algorithm 2

The algorithm can correctly classify whether each review is positive or negative (or both). It can also identify the meal (lunch or dinner) type the reviewer is talking about. It finally gives a list of positive and negative reviews about each dish.

```
*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
MOST LIKABLE COMBINATIONS
*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
['chicken', 'rice']: has occured 7 times
['chowmin', 'pork']: has occured 5 times
['fish', 'rice']: has occured 4 times
```

Fig –4: Output of Apriori Algorithm

The algorithm firstly considers only positive reviews and then outputs the most recommended and suitable combinations.

V. CONCLUSION AND FUTURE WORK

The algorithms are helpful in classifying food items from food reviews. They can classify whether the review is positive or negative or both (for two different items), and based on this classification, they are able to suggest suitable combinations of food after using the Apriori algorithm.

The dataset is very concise. The paper is mainly concerned about the modification in the algorithm, rather than the results (successful on a small dataset). In the future, large datasets can be obtained, and these algorithms can be applied. With some changes to the data in the algorithm, it is supposed to work fine for larger datasets. The algorithms can be used on other types of datasets, such as supermarkets, online stores, etc.

REFERENCES

1. Sreyam Dasgupta, Ronit Chaudhuri, Swarnalatha Purushotham, “Feature Selection for Breast Cancer Detection using Machine Learning Algorithms”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, pg. 2080-2083, July 2019
2. Shiliang Sun, Chen Luo, Junyu Chen, “A Review of Natural Language Processing Techniques for Opinion Mining Systems”, Information Fusion. 36, 2016
3. S.Prakhash, A.Nazick, R. Panchendrarajan, M.Brunthavan, S.Ranathunga, A.Pemasiri, “Categorizing Food Names in Restaurant Reviews”, Moratuwa Engineering Research Conference (MERCon), 2016
4. Pauliina Leppanen, “The Customer Satisfaction Based on the Review Data as given by the Customers”, Bachelor’s thesis, International Business Financial Management, Tampere University of Applied Sciences, November 2016
5. Piero Conforti, Klaus Grünberger, Nathalie Troubat, “The Impact of Survey Characteristics on the Measurement of Food Consumption within Households”, Food Policy, Elsevier, vol. 72(C), pages 43-52, 2017
6. S. SeshathriAathithyan, M. V. Sriram, S. Prasanna, R. Venkatesan, “Hierarchical Classification of Text – An Approach Using NLP Toolkit”, 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 18-19 March 2016
7. Mahesh Kini M, Saroja Devi H, Prashant G Desai, Niranjan Chiplunkar, “Text Mining Approach to Classify Technical Research Documents using Naïve Bayes”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, pg 386 – 391, July 2015
8. R. Premlatha, S. Srinivasan, “Text processing in information retrieval system using vector space model”, International Conference on Information Communication and Embedded Systems, 27-28th Feb, 2014

AUTHORS PROFILE

Sreyam Dasgupta has completerd B-Tech from VIT, Vellore. He is going to study Masters in Data Science at University of Glasgow. He has published four papers –
“Feature Selection for Breast Cancer Detection using Machine Learning Algorithms”, “Extended AES Algorithm with Custom Encryption for Government-level Classified Messages”, “Image Compression using Bayesian Fourier” and “Smart Garbage Monitoring System”.

Suparna Karmakar is currently studying Master Of Computer Applications (MCA) at Cambridge Institute of Technology, Bangalore (permanently affiliated to VTU). She **has** completed her graduation in B.Sc Mathematics Honours from University Of Calcutta. She received certification for securing first place in first year examination of MCA held during the academic year 2018-2019. She was awarded a Junior Scientist certificate from DST-JBNSTS Kolkata, 2014.