International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

CLASSIFICATION OF FOOD RECIPE COMMENTS USING NAIVE BAYES

M. Ravikanth[1], Ch. Reshma[2], B. Sravani[3]

[1]Assistant professor, Dept. of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Andhra Pradesh, India
[2,3]Bachelor of Technology, Dept. of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Andhra Pradesh, India

---

Abstract - *With the growth of web we can find a number of reviews or opinions on any product in many website. Customer spends a lot of time looking for the right product based on the feedback the knowledgeable people share. So, we've created a model that will classify all user feedback into positive or negative for a recipe and show them in graphical representation form. In this paper we apply sentiment analysis on food comments using naive Bayesian algorithm. We use sentiment analysis for analysing comments or reviews into positive or negative. To analyse we will collect comments from web sites and perform pre-processing using natural language processing and apply naive Bayesian algorithm to find class probability to each unique word. Unique words are identified by using bags of words technique. This model helps users to select best recipe by visualizing graph shown for a recipe without spending much time in analysing them.*

***Key Words*: food reviews, sentiment analysis, naive Bayesian algorithm, bags of words, natural language processing.**

1. INTRODUCTION

As per todays internet world we can find hundreds of reviews for any product. Customers want to select best product. To select best they will analyse opinions of experienced people that is how many of them are saying that the product is good and bad. The time taken to analyse each and every product is very high .Even though there are star rating it may not be trusted and we will not know the reason for the best or worst. There are several food websites with recipes on how to cook. In this websites people share their experience about each recipe after cooking. Some people accept food is tasty and others may not. If there is a model which will automatically analyse user comments based on rating will be very useful to the customers to select best recipe in less time. The opinions or reviews given by user are in natural language which is not understood by the machine. Sentiment analysis is a technique which makes machine to understand the human language.

Sentimental analysis is a process of determining a piece of writing into positive, negative and neutral. Sentimental analysis helps large-scale data analysts collect public opinion, perform market research, track brand and product credibility and appreciate client experience. "Opinion mining" is also known as emotional research. The sentimental research has three different levels of reach. i. document level: Sentimental analysis gets the meaning of a complete document. ii. **Sentence level**: Sentimental analysis obtains the sentiment of a complete single sentence. iii. **Sub-sentence level:** Within a sentence, sentimental analysis gets the feeling of a sub-expression. The methods for classifying sentimental research can be defined as follows i. Machine Learning: This method uses a technique of machine learning and a variety of features to create a classifier that can recognize text that communicates feeling. Deep learning methods are popular nowadays ii. **Lexicon-Based :** This approach uses a variety of polarity score annotated terms to determine the overall evaluation score of a given content. The strongest asset of this methodology is that it needs no training data, while its weakest point is that it does not include a large number of words and expressions in emotion lexicons.iii. **Hybrid:** It is called hybrid, the synthesis of machine learning and lexicon-based methods to tackle sentiment analysis. Although not widely used, this procedure provides more promising results than the aforementioned methods. Some of the supervised machine learning techniques that can be applied are k-nearest neighbours(KNN), naive bayesian, linear regression, vector support (SVM), decision tree Section I includes the introduction of sentiment analysis, Section II contains the related work of sentiment analysis of food recipe comments, Section III contains some of the interventions of how the model works, Section IV contains the results of the model we have developed, Section V describes conclusion and future scope of the project.

2. RELATED WORK

Sentiment analysis can be implemented through lexicon based, machine learning, hybrid based method. In lexicon based is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text. With this approach a dictionary of positive and negative terms is needed, with each word being given a positive or negative sentiment meaning. Different approaches to creating dictionaries have been proposed, including manual and automatic approaches. The approach to machine learning makes use of supervised learning techniques. Supervised learning uses labeled data to rate test data positively or negatively. The combination of both lexicon and machine learning approach is hybrid based method. Hybrid method gives more accurate results.

Anshuman, shivani rao and misha kakkar [1]have used sentiment analysis to sort the recipes when an ingredient name is given as input. To sort the recipe they have used sentiment analysis of lexicon-based method. The reviews for number of recipes from various different sites were fetched out and through lexicon-based approach they were analysed . A bag of positive and negative words were used to rate the reviews based on word score comparision. Reviews that has highest score was ranked first position and so on. Pakawan pugsee[2] has done lexicon based sentiment analysis on food recipe comments. In this paper they have classified food recipe comments from a community into positive, negative and neutral. Classification is done by identifying the polarity words from the sentence and by calculating polarity score. Using this method the accuracy score for positive comments is 90% and for negative 70%. Sasikala and Mary inmaculate sheela[3] has done sentiment analysis using lexicon based method on food reviews based on customer rating. They have implemented it using r programming. The opinion word or polarity word from the sentence they have performed pre-processing. All the opinion words and its count are represented in matrix format. Any machine learning algorithm can be used to get the expected result. Kavya suppala and narasinga rao[4] has used sentiment analysis of naïve bayes classifier on tweet data to compare between different tweets. In this they have collected tweets of previous data to train the model and using this labelled data they have predicted test data.

3. METHODOLOGY

Using machine learning methodology, classification of commentary on food recipe using probabilistic model is implemented. Machine learning can be divided into three types i. supervised learning, ii. Unsupervised learning, iii. Reinforcement learning. Probabilistic classifier is one of the supervised learning method. Supervised learning is the one that is directed by an instructor where you can find learning. We have a dataset acting as a teacher and their job is to train the model or computer. Once the model is educated, it can begin to predict or determine when it receives new data. Under probabilistic classifier there are naive Bayes, Bayesian network and maximum entropy algorithms. Among them we are using naïve bayes algorithm.

To develop any supervised learning model first we have to collect previous data. On this data we perform pre-processing to remove duplicates and noise in the data. This data is divided into train and test .On training data we perform naive bayes classifier to get labelled data. Using this labelled data we predict test data. To develop the model we have mainly three steps.

i. pre-processing

ii. Training and testing

iii. Prediction.

The below diagram shows the architecture flow of the system .

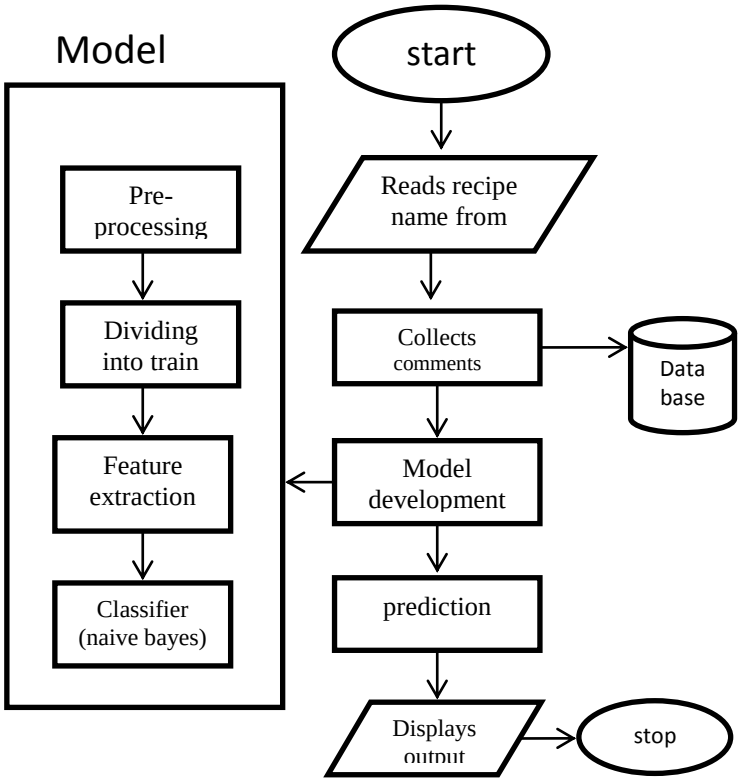

FIG 1: Model flow

i.pre-processing: Preprocessing is a tool used to convert the raw data to a clean set of data. In other words, it is collected in raw format whenever the data is collected from different sources which is not feasible for the study. Pre-processing include following: Removal of duplicates, converting into lowercase letters, removal of stop words, tokenization and stemming. The data collected includes repetitive data that will reduce model accuracy. Therefore, we need to get rid of duplicates. Text often has a range of capitalization representing the start of sentences, proper focus on the nouns. Most of the words in a given text connect sections of a sentence instead of showing subjects, objects or intent. Delete words like "the" or "and" called stop words. Tokenisation is simply a job of chopping a character into bits, called as a token, and at the same time throwing away other characters, like punctuation. Stemming is a mechanism where words are reduced to a root by increasing inflection, usually a suffix, by dropping unnecessary characters. The findings can be used to define commonalities and relationships across large datasets.

ii.Training and Testing:

Before we perform training we need to convert text into vectors called feature extraction. We cannot perform naïve classification of bayes directly on text so there is a need for extraction of features.

Feature Extraction: Once data pre-processing has done it can be used for labelling the data. To label the data we have to extract features from the text i.e., converting text into numbers. Machine learning algorithms can't directly operate with raw text, the text has to be translated into numbers. Specifically, vectors of numbers. For this we are using bags of words technique. A bag-of-words model, or BoW for short, is a way to extract features from text to be used in modeling, for example with algorithms for machine learning. The technique is very simple and versatile, and can be used to remove features from documents in a myriad of ways. A bag-of-words is a text representation which describes the occurrence of words inside a document. It involves two things: A vocabulary of known words, a measure or count of the presence of known words.

Training: Now we're using classifier of naive bayes. Naive Bayes classifiers are a series of Bayes 'Theorem-based classification algorithms. In this it will find the probability of each and every word that are identified in feature extraction. Now, we need to create labeled data. For this, we find the probability of each word for all possible values of the class variable (i.e., positive, negative, neutral).

Testing: In testing we will see how accurate our model is working. We will find f1 score, precision, recall, micro, macro average and confusion matrix. We will also observe the roc curve plot for the model.

Prediction: Once the model has developed we can now predict new data sets. In prediction we will find the probability value of all classes(i.e., positive or negative) for a comment. Among the class probability values we will assign the class which has highest value.

Now, we're going to see how the labeled data is created.

I. We contain a training data set containing documents belongs to classes say class positive (pos) and negative (neg).
II. Now find the probability of both class p(pos) and p(neg). p(pos)= number of positive documents / total number of documents.
III. Now identify word frequency of class positive and negative. For example consider the word tasty. In the entire document count how many times the word tasty has occurred in positive document and negative document).
IV. Calculate the probability of keywords occurred for each class.
P(word1/pos) = word frequency1 +1/total number of word frequency of class pos
P(word1/neg) = word frequency1 +1/total number of word frequency of class neg
P(word2/pos) = word frequency1 +1/total number of word frequency of class pos
P(word2/neg) = word frequency1 +1/total number of word frequency of class neg
……..and so on.
P(wordn/pos) = word frequency n +1/total number of word frequency of class pos
P(wordn/neg) = word frequency n +1/total number of word frequency of class neg
V. New document N is classified based on probability value of class positive and negative.
   a. P(pos/N) = p(pos) * p(word1/pos)*p(word2/pos)……*p(wordn/pos)
   b. P(neg/N) = p(neg)*p(word1/neg)*p(word2/neg)…….*p(wordn/neg)
VI. After calculating probability for both class pos and neg the class with higher probability is assigned to new document N.

4. RESULTS AND DISCUSSIONS

In this experiment we have used amazon fine food reviews to train the model. It contains 5,68,454 reviews for 74,258 recipes. For experiment purpose we have used 2,00,000 comments for 3000 recipes. We have created a user interface where user select a recipe name for which we wants to observe the analysis of reviews. After submitting they can observe a pie graph with number of positive and negative comments for that recipe.

FIG 2: Input

The above fig is the user interface where he selects a recipe name from the list.

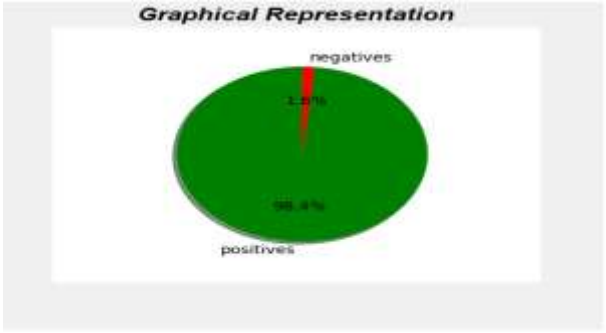

FIG 3: Output

The above graph shows the positive(green color) and negative(red) comments for the selected recipe.

Out[26]:

	Pos_Words	Pos_Importance	Neg_Words	Neg_Importance
0	like	-4.406478	tast	-4.126758
1	tast	-4.412099	like	-4.246824
2	love	-4.433405	product	-4.390704
3	great	-4.458662	one	-4.732713
4	good	-4.476377	flavor	-4.764328
5	flavor	-4.598010	would	-4.856262
6	use	-4.708053	tri	-4.901650
7	product	-4.714912	buy	-5.005655
8	one	-4.797766	good	-5.008839
9	tri	-4.882972	coffe	-5.043479
10	coffe	-4.922962	order	-5.072395
11	tea	-4.969096	use	-5.102606
12	make	-5.025104	get	-5.134089
13	get	-5.080651	dont	-5.235991
14	buy	-5.213327	box	-5.302261
15	price	-5.286251	tea	-5.373534
16	best	-5.289779	food	-5.388235
17	time	-5.296074	even	-5.390599
18	food	-5.301066	amazon	-5.484577
19	realli	-5.330851	eat	-5.505956

FIG 4 : Top 20 features table

Above table is the top 20 positive and negative words and their corresponding probability values.

The model we have developed has shown an accuracy of 93% .

```
AUC Score 0.9356708485671656
macro f1 score for data : 0.6150328604181582
micro f1 scoore for data: 0.8686633137845223
hamming loss for data: 0.13133668621547778
Precision recall report for data:
              precision    recall  f1-score   support

           0       0.92      0.18      0.30     18099
           1       0.87      1.00      0.93     96827

   micro avg       0.87      0.87      0.87    114926
   macro avg       0.90      0.59      0.62    114926
weighted avg       0.88      0.87      0.83    114926
```

FIG 5: Measures of the model

Above figure shows the accuracy, precision, recall, f1-score for positive and negative comments.

5. CONCLUSION AND FUTURE WORK

In conclusion, we have developed a model which performs sentiment analysis on amazon fine food reviews using machine learning. For processing and analysis human language we have used Natural language processing tool kit on dataset . Bags of words technique is used to extract features from the text. Classification was done using naive bayes by calculating the probability of new data and assigning class which has highest probability value. The model developed has highest accuracy and we can use even effective methods. For further work we can have personal profiles of people who comments on recipe like age and gender so that we can analyse which age group people have liked or disliked the recipe.

REFERENCES

[1] https://www.researchgate.net/publication/317418948_A_rating_approach_based_on_sentiment_analysis.
[2] https://ph01.tci-thaijo.org/index.php/ecticit/article/view/54421/45192.
[3] https://acadpubl.eu/hub/2018-119-15/2/373.pdf.
[4] https://www.ijitee.org/wp-content/uploads/papers/v8i8/H6330068819.pdf.
[5] https://arxiv.org/ftp/arxiv/papers/1612/1612.01556.pdf.
[6] https://pdfs.semanticscholar.org/ccbf/5b65c00e666309 3465f3e784c8b649dbeb32d.pdf.
[7] https://www.researchgate.net/publication/312176414_Sentiment_Analysis_in_Python_using_NLTK.
[8] https://www.researchgate.net/publication/329513844_A_Real-Time_Aspect-Based_Sentiment_Analysis_System_of_YouTube_Cooking_Recipes.