IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 3, Ser. V (May - June 2019), PP 59-64
www.iosrjournals.org

Online Product Review Classification

Fattesingh Rane[1], Gaurish Kauthankar[2], Akhil Naik[3], Sulaxan Gawas[4], Kedar Sawant[5], Rajesh Gauns[6]

[1]*Agnel Institute of Technology and Design AICTE, Goa-University Bicholim-Goa, India*
[2]*Agnel Institute of Technology and Design AICTE, Goa-University Bicholim-Goa, India*
[3]*Agnel Institute of Technology and Design AICTE, Goa-University Mapusa-Goa, India*
[4]*Agnel Institute of Technology and Design AICTE, Goa-University Bicholim-Goa, India*
[5]*Agnel Institute of Technology and Design AICTE, Goa-University Sanquelim-Goa, India*
[6]*Agnel Institute of Technology and Design AICTE, Goa-University Ponda-Goa, India*

Abstract: *Reviews are the most important part that people look upon while purchasing a product online. The problem in the existing system is sometimes the user review and the rating mismatch each other. This happens when the user forgets to update either review or rating when the user updates the review or while providing a new rating for the product, the user might randomly put some wrong rating or undesired rating.*
The main aim of this project is to tell the user whether a product is good or bad based on the reviews provided by other users and to provide a better rating for the product by analysing sentiment. To classify the reviews into good or bad, the system uses two machine learning algorithms KNN and Naïve Bayes classification algorithms and to stem the review porter stemmer algorithm is used and to compute new rating system uses rule-based extraction method. K Nearest Neighbour will select the nearest neighbour class to the test review and classify the review into two classes that are either class = good or class = bad, whereas the Naïve Bayes algorithm uses a probabilistic approach to classify the product into good or bad by selecting the highest probability class label.

Keywords: *review, rating, classification, stemming, probability, KNN, Naïve Bayesian,*

Date of Submission: 12-06-2019 Date of acceptance: 28-06-2019

I. Introduction

Reviews are provided by the customer after using the product which provides us with the disadvantages and advantages of that product helping other customers to make the correct decision of the product. The review helps other customers to know the detailed experience of the product. Reviews are written in the English language with content specifying the advantages, disadvantages, features, faults, etc. reviews also differ from product category such as electronic devices products will provide different information whereas non-electronic devices will provide different information.

Rating is the overall score of the product given by the customer. Rating helps other customers to know the overall experience of the product experienced by that customer. This might be more useful since most people don't like reading and they can simply refer to the score and make the correct decision about the product before buying it.

As the rating have an advantage over reviews that it reduces the burden of reading a big paragraph written by the customers with just a single number but the rating has a drawback that it provides less information to the buyer since review share detailed information and experience of the customers who have worked with that product.

1.1 Problems with reviews and ratings

The problem in the existing system is sometimes, the user review and the rating mismatch each other. Example consider the initial review as 'calculator is good' and 'rating is 4' but after using the product user changed his review about the product as 'calculator is bad' but he has forgotten to change the rating.

This happens when the user forgets to update either review or rating whenever the user updates his existing review or while providing a new rating for the product, the user might randomly put some wrong rating or undesired rating. If some user is willing to buy the product and if he simply checks the rating than that user might get incorrect information about the product as we can see the new review says 'calculator is bad' but it was still 'rating as 4' making the users consider the product is good based on rating information.

1.2 Solution approach

The main aim of this project is to tell the user whether a product is good or bad based on the reviews provided by other users and to provide a better rating for the product by analysing sentiment. Reviews contain different types of words like noise words such as the, a, you, etc. then there are important words that provide information about the review such words are good, better, best, worst, awesome, bad, disgusting, etc. such words gives two classes of the review's positive classification and negative classification. Also, there are some special words like but, then, as, compared, etc, which combines two different meaning sentences into a single sentence. Example product A has better specifications than product B.

To classify the reviews into a good or bad rating, the system uses two machine learning algorithms KNN and Naïve Bayes classification algorithms. KNN will select the nearest neighbour class to the test review and classify the review into either good or bad, whereas the Naïve Bayes algorithm uses a probabilistic approach to classify the product into good or bad by selecting the highest probability class label. Both the algorithms nearest neighbour and naive Bayesian are going to generate their own dataset based on the training data provided to them. The dataset generated by these algorithms is specific to them that is you cannot use generated data of Naïve Bayes for KNN algorithm and vice versa.

To stem the review, the Porter Stemming algorithm is used which removes the suffixes of the words and root words or stem words are kept in the review. Example if there exists an incorrect word like 'goods' the algorithm will remove the suffix 's' from 'goods' and it will keep the stem word which is 'good' in the review.
Rating for the product will be computed by using rule-based extraction method. This method checks the occurrence of positive and negative words and maps these words in the dataset and fetches the appropriate score and computes the rating. This method also takes care of the tenses.

II. Literature Review

Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier paper specifies that Naïve Bayes is the probabilistic classifier algorithm based on Bayes theorem considering strong independence assumption and it is a baseline for text categorization and document classification belonging to a particular category based on the frequency of words and attributes. Naïve Bayes needs less training data and it uses conditional probability model. Bayesian classification is a learning algorithm that computes the prior knowledge of the data and combines with newly computed data. In Naïve Bayes entire document classification is made.
Paper uses the movie and hotel review dataset and divides it into two files pos.txt and neg.txt.

These two files are read and two empty lists are created and each review is assigned with pos tag and neg tag and stored into this empty list. Now this list is fed with training data 3/4th sentences of this are kept for training and rest 1/4th for testing.
Authors method of training data creation is shown below

2.2 KNN Algorithm [3]

Sentiment mining is the most significant data mining where the required data can be retrieved based on the features of the collected data. KNN is a lazy learning approach. The objects are classified depending on the count of occurrence and are assigned with the class having the smallest Euclidean distance among its nearest neighbour. This method keeps entire training data for learning. The Nearest neighbour selects the smallest neighbour from entire computed values.

In this proposed system Dataset of movie reviews is considered which is used to classify a particular movie into negative or positive by using K-NN. The dataset is distributed into two separate files i.e. positive words and negative words which are stored separately and also their corresponding count. Now based on the test data the positive and negative words are matched and their corresponding counts are squared and added and finally square-rooted. This result is checked for both positive as well as negative words and the class having the minimum value will be assigned as a final result.

2.3 Porter Stemmer [7]

M.F. Porter proposed the Porter Stemmer stemming algorithm which removes the suffixes from the given words which are used for the field of data retrieval. There are definitions or production rules for removing the suffixes of the words. The main rule for removing suffixes is (condition)S1-->S2 where S1 is the suffix to be removed and S2 is the replacement for the suffix S1. In this project, some of the definitions or the production rules for removing the suffixes are changed or removed as those rules alter the meaning or made those words incorrect which were actually required by the system in the later stages.

2.4 Feature Extraction Method [4]

Harsh Chheda, Ankit Patel, Ashutosh Bhatt, Kiran Gawande proposed a system sentiment analysis and amazon review classification system where the system basically processes the reviews about a product and tells whether the given product is bad or good.

The method that has been followed in the system is that the system will remove all the special words from the review such as (. , % & $?etc)and the cleaned data will be further given to the next stage.

In the 2nd stage, the sentiments from the reviews are extracted by using the post tagging and rule-based extraction method. After the sentiments are extracted, they are sent to the polariser which returns 1 if the sentiment is positive or returns -1 if the sentiment is negative. Once the entire review is polarised a collective polarity of the review is calculated. If the collective polarity is less than 0 then the outcome is negative and if the collective polarity is greater than 0 then the outcome is positive. Using this method, the system basically tells the user if the given product is good or bad.

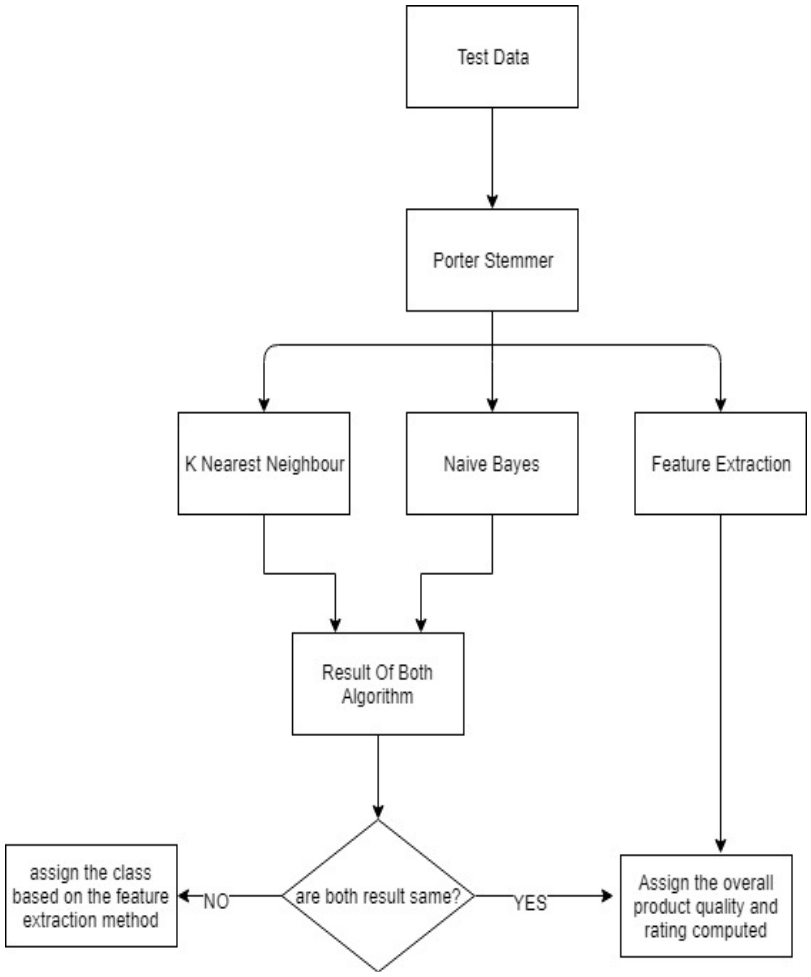
2.5 Suffix Stripping [9]

Suffix stripping is basically automatically removal of the suffix from the English words. The algorithm for suffix stripping is implemented using BCPL. It effectively works by executing complex suffixes as a group of simple suffixes in steps.

2.6 Opinion mining and Sentiment analysis [10]

Opinions are very important when it comes to making a decision to choose from different options. whereas sentiment analysis basically deals with polarity and emotion detection. Both these fields are used in data mining and NLP methods. Currently, these techniques depend on vector extraction. Sentiment analysis can be done the methods such as keyword spotting, lexical affinity, statistical methods, etc. The new technique used in sentiment analysis is a multimodal analysis which deals with audio or audio-visual format rather than text

System Flow Diagram



III. Dataset Description

For any software to be built, there is a need for effective and accurate data and for a learning algorithm, the incorrect data cannot be given as part of the training. The product review dataset is available on two websites Kaggle.com and data.world.com

Both these data sets are equally accurate. The dataset from data. the world contains 10 attributes out of which only 4 are required that is ID, Class, Review text and Rating whereas the data from Kaggle.com contains only two attributes Class and Review text.

3.2 Attribute and Class label Description

ID Unique Identifier for that product name 1

Class Category that marks whether the product is good or bad

Review text Comment or customers feedback provided to the product

Rating provided to that review.

IV. Classification And Stemming

A Introduction

Classification is the technique used in data mining for classifying the attribute or the data into a specific class or category by means of predictions. Predictions are done based on probability or based on the closeness of the data points to that particular class. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

In classification, the most important part is feeding the correct data to the learning model because feeding incorrect data will lead to incorrect classification. The model of classification finds the relationship between the test data and the trained data by prediction and assigns the test data a particular class. Trained data is the data whose class is already known and test data is data whose class is unknown and need to be predicted by means of prediction.

Classification uses the supervised learning approach. In supervised learning, the mapping of data and class of the training data is already known. And the goal of classification is to predict the class of unknown data points. Data set for any type of supervised learning classifier is divided into two categories training dataset for building model and testing dataset for testing the model which is being built.

B Naïve Bayes Classification Algorithm [6]

Naïve Bayes classification algorithm is based on Bayes theorem claims the probability of occurrence of an attribute, based on previous information of conditions that might be related to the attribute. Naïve Bayes makes the assumption that the value of the particular attribute does not depend on the values of another attribute, of given class label. Naïve Bayes is the most famous algorithm for text categorization and classification.

Naïve Bayesian classification technique is dependent on Bayes theorem and the Bayes theorem is stated below. Consider X to be an attribute whose class is not known and say H be the probability that attribute X belongs to that class C. The Bayes theorem can be used to calculate the posterior probability as

$$P(C/X) = \frac{P(X/C) * P(C)}{P(X)}$$

- *P(C|X) be conditional-probability of occurrence of C given that event X has already occurred (need to be predicted)*
- *P(C) be the probability of event C to be true*
- *P(X|C) be conditional-probability of occurrence of X given that event C has already occurred*
- *P(X) be the probability where event X is true*

1) Algorithm steps

1. Consider training dataset D that is mapped to the class label. Each element is represented by a vector,
2. Assuming there are m classes. if we need to classify an element X whose class is not known than the classifier will estimate that the attribute X belongs to a particular class with higher future probability, conditioned on X. i.e., the Naive Bayes method assigns an unknown element X to the particular class if true above future probabilities are computed using Bayes Theorem.

Naive Bayes is fast and easy to estimate class for test data. It also performs well in multiple class prediction. When consideration of non-dependency holds, a Naive Bayes method performs better as compared to the KNN method. Naive Bayes is faster than KNN and is successful in producing the desired inference on huge data. Naive Bayes deals with noisy data and takes less memory compared to KNN. Representation is also easy to understand with Naive Bayes than KNN.

2) Algorithm: Naïve Bayesian Classifier

Learning: Dataset for training the classifier with a correct mapping of class label to its values
Input: A sample of data for testing the classifier with the unclassified class label
Output: Test sample mapped to the correct class label
Steps
1. Train the classifier using the following steps
a. Transform the data into a frequency table
b. Find probabilities and compute a Likelihood table
2. Give the test data for the classifier.
3. Compute the future probability of each class.

B. 4.3 K Nearest Neighbour Classification Algorithm [3]

The KNN Algorithm is the easiest of all the ML algorithms. It is based on the principle that the samples that are similar, generally lies in close vicinity. KNN is an instance-based learning method. Instance classifier is also known as lazy learners because they store the training data and until a new, unlabeled sample is to be classified it doesn't build a classifier.

A case is classified based on a maximum number of votes of its neighbors and the case being given to the class that is most common amongst its KNN measured by a distance function. If K= 1, then this case is given to the class of its nearest neighbor.
Choosing the best value for K is done by first checking the data. In general, a higher K value is more accurate as it decrements the whole noise but there is no guarantee. Cross-checking is another way to reflectively determine a good K value by using independent data to validate the K value. Historically, the best K for most data follows in the range of 3-10. That generates a better outcome than 1NN.
Input: k nearest training data
Output: Class membership is an object which is classified by a higher count of a vote of its neighbors
Algorithm
1. Determine parameter K= number of nearest neighbors.
2. Calculate the distance between the query-instance and all the training samples.
3. Sort the distance and determine nearest neighbors based on the K-to minimum distance.
4. Gather the category of the nearest neighbors.
5. Use a simple majority of the category of nearest neighbors as the prediction value of the query instance.

4.4 Feature Extraction Method [4]

In feature extraction method we extract the word these words represent features of that review which has been written by concerned user. Extracting features means fetching the words from the line of text that marks the positive qualities and negative qualities of the product example of feature words are good, better, best, worst etc.
Simply extracting features as an individual word might lead to incorrect inference so the features extracted must be based on following cases
1. The word fetch from i th location must be checked for two conditions based on type of iteration you are performing. Basically, there are two cases possible which are
a) Fetched word can be negated by its preceding word in forward or reverse search so it is necessary to take care in reverse search for negation condition since the match would be for the word and not for the negator.
b) Fetch word can also be preceded by a word that increases the score of that word e.g. if word match is good with score 4 and if proceeding word is very then the score of word will be incremented by 1 letting total score very good as 5.

2. Sentences differ their meaning by time for example whatever was good yesterday might not be good today same way whatever will be good in the future is not good in current date. So, we must also take care about the tense of the sentence before rating the words.
3. There are sentences that leads to two dependent conclusions when read but we can split such dependency-based sentences to get the overall score such sentences are
separated by words like that, than, as etc.
4. The sentences can also be comparative as specified above and to process such sentences we need to handle them separately or independently of the main searching process in the special cases handler
1) Special cases possible if and only if the sentence contains the name of the product which might compare two products.
2) Special cases also possible due to presence of special words but this is rare case example of such sentence is "its better to buy some other product rather than buying this one"

V. Conclusion

Product review can be classified by using data mining techniques with the use of two machine learning algorithm KNN and Naïve Bayes. KNN finds the minimum Euclidean distance which is a lazy learning technique. Naïve Bayes uses a Bayes theorem and probabilistic model to classify the reviews. We have the KNN and the Naïve Bayes model to classify the reviews into either good or bad whereas to compute a new rating we extract the features from the sentences and extract the words which mark the positive features and negative features of the review.to process any review we must first stem the sentences to compute new rating before any of the methods can be applied. Stemming does the job of removal of stems in the words and also removes the special symbols finally using all these models Porter Stemmer, Naïve Bayes, K-Nearest Neighbor and Feature extraction technique we can classify the overall quality of the product and also assign a new rating between the range of 1-5.

Earlier people used to manually go to the reviews and check all the reviews manually when they wish to purchase any product from E-shopping sites. This process is time-consuming as well as it does not help the user to get all the necessary information what he actually wants. The proposed system reduces the effort of manual searching of reviews by directly classifying the reviews about that product and thus finally telling whether the product is good or bad.

The future scope for this system will be instead of simply telling the product is good or bad, it can also classify the product into different rating classes like Excellent, Good, Average, Bad, Worst. The optimization of feature extraction can help in better performance and accuracy.

Acknowledgment

We are greatly indebted to our Principal, Prof V Mariappan, Prof. Snehal Bhogan Head of Computer Engineering Department, our Guide Prof. Kedar Sawant, Agnel Institute of Technology and Design, Assagao Goa, who gave us the opportunity to do the project on the topic "**Online Product Review Classification**" and also for their valuable guidance throughout the dissertation, without which the study undertaken would not have been accomplished.

Our sincere thanks to our Co-Guide, Prof. Rajesh Gauns, Faculty and Staff, Department of Computer Engineering, Agnel Institute of Technology and Design, Assagao Goa and our colleagues for their constant support and encouragement rendered throughout.

References

[1] Pang-Nin Tan, Michael Steinbach and Vipin Kumar "Introduction to Data Mining by Pearson Education" ISBN:9780321321367, Indian Edition 2014

[2] Jiawei Han, Micheline Kamber and Jian Pei "Data Mining Concepts and Techniques", by Moran Kaufmann Publishers, ISBN978-1-12-381479-1, Third Edition 2012

[3]. Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, and Sweta Tiwari "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier", Mon, 31 Oct 2016

[4]. Aashutosh Bhatt, Ankit Patel, Harsh Chheda, and Kiran Gawande "Amazon Review Classification and Sentiment Analysis", 2015

[5] Bayes theorem
en.wikipedia.org/wiki/Bayes%27_theorem

[6] Naïve Bayes initial approach
www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[7] Porter Stemmer Algorithm
tartarus.org/martin/PorterStemmer/
www.youtube.com/watch?v=GQ1sXx8hH4k

[8] Text classification technique and algorithms
www.nltk.org/book/ch06.html

[9] "An algorithm for suffix stripping" M.F. Porter (Computer Laboratory, Corn Exchange Street, Cambridge), 2006

[10] "New Avenues in Opinion Mining and Sentiment Analysis" Erik Cambria, National University of Singapore, Björn Schuller, Technical University of Munich, Yunqing Xia, Tsinghua University, Catherine Havasi, Massachusetts Institute of Technology, 2013

Fattesingh Rane. "Online Product Review Classification." IOSR Journal of Computer Engineering (IOSR-JCE) 21.3 (2019): 59-64.