# Warehouse Review Classification Using Naïve Bayes Classifier

Sinchana T N
M.Tech (CSE)
Nitte Meenakshi Institute of Technology and Management
Bengaluru, Karnataka
e-mail: *sinchanatngowda@gmail.com*

Dr. Jharna Majumdar
Dean R&D, Prof & Head of M.Tech (CSE)
Nitte Meenakshi Institute of Technology and Management
Bengaluru, Karnataka
e-mail: *jharna.majumdar@gmail.com*

***Abstract***— A customer review is a review of a product or service made by a customer who has purchased the product or service. Customer reviews are a form of customer feedback on electronic commerce and online shopping sites. In this study, a review of a warehouse by a customer who has booked a warehouse for storing their goods will be classified. By this classification a customer will come to know whether the warehouse is suitable for storing their goods or not as the proposed model will classify the reviews into positive and negative reviews.. This paper proposes an approach to perform subjectivity classification on feedback text based on a supervised machine learning algorithm, Naive Bayes. Experiment studies have been conducted on warehouse reviews. The results show that the performances of the proposed approach are comparable to those of the existing english subjectivity classification studies.

***Keywords***- *warehouse review, Naïve Bayes.*

*****

## I. INTRODUCTION

In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

In our case study on analyzing sentiment, you will create models that predict a class (positive/negative sentiment) from input features (text of the reviews, user profile information,...). In supply chain management ,warehouse management and storage management inside the warehouse is an important stage. So In this study I am trying to book a warehouse based on the reviews about those warehouses made by the customers who used those warehouses for storing their goods. One can book the warehouse based on the percentage of positive reviews a warehouse got by the customers.

So In order to classify the reviews made by the customers I have used the naïve bay's classification algorithm for warehouse booking into a positive and negative reviews, So that a user can book the warehouse which has got more positive reviews.

## II. LITERATURE SURVEY

### *A. Sentiment Analysis*

It is also called as opinion mining which will test the opinions or emotions made by the users about any product by giving the comment or feedback, reviews etc .Opinions means statements made by the user which tells one's emotion or sentiment towards that product. Because of the improved technology and the internet selling a product or marketing a product has been reached to a higher level. As a result many companies will decide before launching a product depending on the reviews made by the customers. If the companies wants to know about how their product has been received by the market they should look into the reviews and these reviews are playing an important role on companies. But the problem here is thousands of reviews are generating for a single product and companies can't process all these reviews to get the opinion of an user about that product. In order to build a domain knowledge for a system it is very necessary of expert assistance and there by we can make the system to learn about the domain specific words. The proposed system will extract the opinions made by an user and these extracts contains the opinion words ,polarity in the weight format and also contains the type of feature for which this belongs to.

### *B. Feature Extraction*

This is one of the important task to be taken before analyzing the opinions of the users.

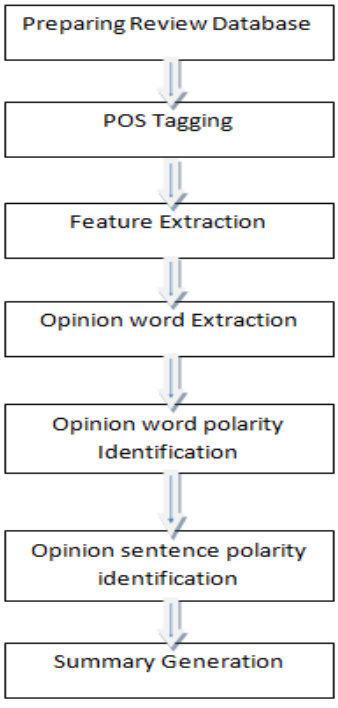



Fig-1: Basic steps of feature extraction

The above figure shows the generic model for the feature extraction which are extracted from the opinion information. Here the first step is creating an information database and next tagging of pos will takes place on the review, next will be the feature extraction which uses a grammar rule Which is as shown below adjective + noun, and so on here nouns represents the features and adjectives represents the sentiment words. Finally we will extract the opinion words followed by its polarity ID.

There are many algorithms available in opinion mining in order to classify a review as positive or negative. These algorithms can be apply to different kinds of data like movie, reviews of products.

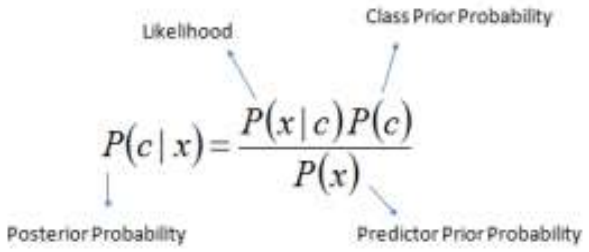
In thousands of reviews user can't find the positive reviews about a product by seeing each and every comments, So we are using different machine learning algorithms to classify them by the opinion of the users about a product.

## III. NAÏVE BAYES CLASSIFIER

Based on the Bayes Theorem an efficient classification technique has been developed and it is called as Naïve Bayes Classifier which assumes that all the predictors are independent of each other. In other words, there is no relationship between the presence of a particular feature in a class with any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

When there is a large dataset we can use this Naïve Bayes model because it is easy to build. It is been called as highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:




$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

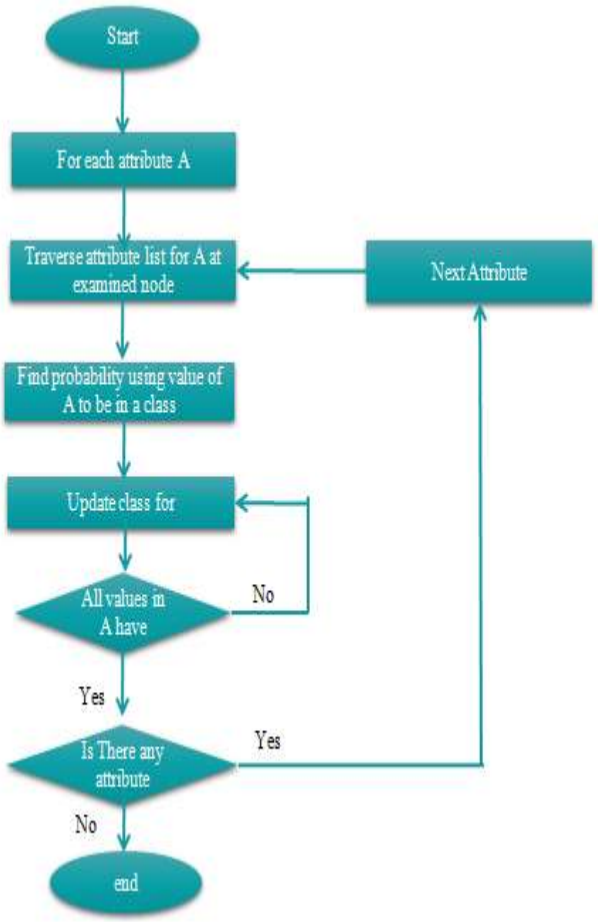



Fig-2: Naïve bayes Classifier

Set Up

Here we are taking the data from user reviews on warehouses. We'll be classifying positive reviews against negative reviews.

1. Preprocessing: Used 100 reviews for training and 50 reviews for testing.

2. Features: The features we're going to use are simply the lowercased version of all the words in the review. This means, in order to get a list of these words from the block of text, we remove punctuation, lowercase every word, split on spaces, and then remove words that are in the NLTK corpus of stopwords (basically boring words that don't have any information about class).

**Naïve Bayes**

*A. Counting the words:*

- Step 1 is telling about calculating the count of words in reviews obtained by the user feedback which belongs to each class, it may be positive class or negative class.
- Now we are having the count of words for each class i.e positive and negative and we will create two training files. This is in the form key value pair where key represents the word in a file and value represents how many times that word has occurred in the review
- Secondly we have to calculate the probabilities of that respective class.
- It is very difficult to getting this probability values for each class because calculating this probability is only

 ISSN: 2321-8169
12 – 15

based on the previous reviews and there may be a change in that number also.

- Finally we have to store this counter dictionary into an variable.

*A. Keeping a Record:*

For the purpose of keeping the records we should consider few things. Keeping the correct records and incorrect records and we will come to know the percentage of correct records.

*B. Algorithm time:*

First we need to clean the data in a give review document using some cleaning method also use some tokenization methods. In the given text document we will calculate the probability of each class. In Naïve Baye's that value will be considered as follows.

```
P(class) * P(word1| class) * P(word2 | class) * ... * P(word n | class)
```

The meaning of the above mentioned representation is the probability of a word which an user expects to see , given the set of words within that class.

This will tell the number of that word an user see those words in the given document with respect to that class divided by total number of words in that class. Here consider X as the total number of words which are present in a particular class. Y will be the number of words you are looking for in that given document with that respective class. Y/X will give the the probability that a word one user is seeing in that class.

If an user haven't see the word before then there will be a problem. In this class we have to assign 0 to Y, So the value of the probability of a word in that class will be 0. So there is a method called additive smoothing which will make sure that the value of the numerator statement should not be 0.

Finally by applying above method the equation turns into (Y+1) / (X+number of words in vocabulary) where number of words in vocabulary represents the total number of unique words that we have seen from the review document.

## IV. RESULTS AND CONCLUSION

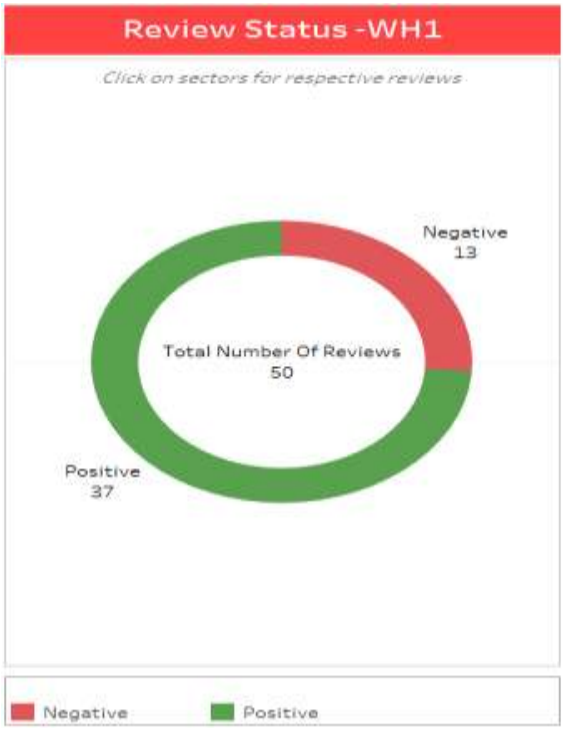



Fig-3: Classified Reviews


Fig-4: Positive reviews

 ISSN: 2321-8169
12 – 15

| Satus |
| --- |
| Negative |
| Bad |
| Difficult in rainy season as leakages in ceiling. |
| Don't want to book again |
| No airconditioning facility |
| Not bad. |
| Not enough machine handling equipments |
| Not good.. |
| Not nice |
| not price worthy |

Fig-5: Negative Reviews

| RESULT | |
| --- | --- |
| Total Number of Reviews | 50 |
| Total Number of Positive Reviews | 37 |
| Total Number of Negative Reviews | 13 |
| Percentage of Positive Reviews | 74.00% |
| Percentage of Negative Reviews | 26.00% |

Fig-6: Final result of review classification

In this study I have trained the model by using training dataset which is containing a 100 of reviews made by the customers about a warehouse which is been used by the customers, each review belongs to either positive or negative class. By applying Naïve Bayes Model predictors, test datasets which is having 50 reviews are classified into 37 positive reviews and 13 negative reviews.

## REFERENCES

[1] Bhuvana, Dr.C.Yamini, "survey on classification algorithms for datamining: (comparison and evaluation).", Aug 2015
[2] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande, "Amazon Review Classification and Sentiment Analysis", Computer Department, Sardar Patel Institute of Technology, Andheri –west, Mumbai-400058, India.
[3] Weishu Hu, Zhiguo Gong, Jingzhi Guo, "Mining Product Features from Online Reviews" IEEE International Conference on E-Business Engineering, 2010.
[4] Gurneet Kaur, Abhinash Singla, "Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm " Jan 2016 .
[5] Neethu M S, Rajasree R "Sentiment Analysis in Twitter using Machine Learning Techniques", July 4 - 6, 2013
[6] Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statist. data. Machine Learning.
[7] Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2001). Generalized belief propagation. In Adv. NIPS 13, 689–695.
[8] Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval, 4(2), 133–151.
[9] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, 103–130.
[10] Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. Proc. UAI-98 (pp. 43–52).