Feature Level Text Categorization For Opinion Mining
Gandhi Vaibhav C.
Computer Engineering
Parul Institute Of Engineering & Technology
Gujarat Technological University

Assistant Professor Neha Pandya
Department of Information Technology,
Parul Institute Of Engineering & Technology
Gujarat Technological University
Abstract

Text classification is an important research area as it enables the computers to work intelligently process unstructured data. This unstructured data is a rich source of information for industries, huge organization, etc. Most of such opinion rich data (more than 85\%) is in text format. In this work we have observed the effect of different algorithms on the text data including the Naïve Bayes. Our main focus is on improving the classification of text efficiency with using Naive Bayes Algorithm. with Feature selection sentimental analysis procedure we get the results according to users required attribute or entity.

KEY WORDS: Text Classification, Opinion Mining, Naïve bayes

1. Introduction

An important part of our information-gathering behaviour is always been to check what the other people are thinking about it. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. “What other people think” has always been an important piece of information for most of us during the decision-making process.

The Internet and the Web have now made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics that is, people we have never heard of and that’s why opinion mining is called the voice of the customer. And conversely, more and more people are making their opinions available to strangers via the Internet. The interest that individual users show in online opinions about products and services, and the potential influence such opinions wield, is something that vendors of these items are paying more and more attention to.

In today’s world, there are so much data available on the internet. It includes the customer reviews on different products. It is a general tendency that before we go for purchasing any product, we go thru the reviews written on the website of that product. By reading those reviews, customer takes decision. Sometimes there are so many reviews that the customer is not able to read, for that the opinion mining is used to help the customer.

The reviews of the customers also help the other customer in getting the suggestions or feedback for the developer of the product. By these reviews, the company can come to know that what is lacking in their product. For example, for mobile, it has been written that, the battery life of mobile is very less, or the voice clarity is not good, so the company can make the battery life and voice clarity better in the next model of that product. By the comments or reviews, the company of that product can come to know that, what are the reasons to like the product and what are the reason for not liking the product.

1.1 Types of Opinion Mining

There are three types of opinion mining approach.

[1] Feature level or Phrase level

In this, for the product, the particular features are classified and for those features, the comments or reviews are taken separately.

[2] Sentence level

In this, the comments or reviews are opinionated. The benefit of this approach is in this, the customer can come to know about so many different types of customer’s reviews. In this approach, it mainly differentiate between the subjective and objective information. The subjective information is the opinion, which can be negative or positive and the objective information is the fact.


In this the whole document is written for the product, it is written by only one person. So, it is not as useful because the customer will come to know the review of only one customer.

2. Naïve Bayes Algorithm

This algorithm Called as Naïve Bayes because its based on “Baye’s Rule” and “naively” assumes independence given the label like

- It is only valid to multiply probabilities when the events are independent
- Simplistic assumption in real life
- Despite the name, Naïve works well on actual datasets

The Naïve Bayes classifier, also called simple Bayesian classifier, is essentially a simple BN. Since no structure learning is required, it is very easy to construct and implement a Naïve Bayes classifier. Despite its simplicity, the Naïve Bayes classifier is competitive with other more advanced and sophisticated classifiers.

The Naïve Bayes method is a kind of module classification under the known prior probability and class conditional probability, its basic idea is to calculate the probability that the text belong to. The probability of the class the text
belong to is equal to the composite expression of
the probabilities that lexical terms in the text
belong to,

The steps for preprocessing and classifying a new
document can be summarized as follows.
[1] Remove periods, commas, punctuation, stop
words. Collect words that have occurrence
frequency more than once in the document.
[2] View the frequent words as word sets.
[3] Search for matching word set(s) or its subset
(containing items more than one) in the list of
word sets collected from training data with that of
subset(s) (containing items more than one) of
frequent word set of new document.
[4] Collect the corresponding probability values of
matched word set(s) for each target class.
[5] Calculate the probability values for each target
class from Naïve Bayes categorization theorem.
Following the steps mentioned above, we can
determine the target class of a new document.

The equation of Bayesian classifiers use Bayes
theorem, which says

\[
p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)} \quad ...........[1]
\]

Where \ p(c_j | d) = probability of instance
d(document) being in class c_j,
This is what we are trying to compute
\[p(d | c_j) = probability of generating \]
instance d given class c_j,
We can imagine that being in class c_j, causes you to
have feature d with some probability
\[p(c_j) = probability of occurrence of class \]
c_j.
This is just how frequent the class c_j, is in our
database
\[p(d) = probability of instance \]
d(document) occurring
Naive Bayes is fast, accurate, and can reflect the
influences to the final conclusion that all
attributes produce, and the realization of the
algorithm is relatively simple, only one scan of
the data set, and suitable to online model
construction. Besides it is also a kind of very
strong algorithm, it has rather strong ability in
resisting disturbs, therefore more and more experts
give attentions to it.

Naive Bayes is a kind of probability classification
model based on two assumptions:
[1] It requires all attributes in given categories
takes independent values, which means any
attributes should not depend on other attributes.
[2] The lengths of texts are independent of their
categories. These assumptions is seldom met in
practical applications.

3.Logical Steps For The Opinion Mining
Approach

[1] First of all, generate the files of good words and
bad words.
[2] Rearrange it, in one phrase, two phrase words.
[3] Assign weights or numbers to all the words,
negative number to bad words and positive number
to good words.
[4] Generate training data set, means generate the
some numbers of comments.
[5] Apply the algorithm, and two files on the
training data set.
[6] Finally apply it on the live data.

3.1Expansion of Files of Good and Bad Words

Figure 1: Iterative Process For Updating Files

From this figure good (positive) words and bad
(negative) words will be taken out from the
different websites. And the other remaining words
which is not necessary or not requiring for the
opinion time would be left at that time. At this way
we can generate the different the good words and bad words.

4. Implementation Work Flow Approach

[3] Implement the opinion mining algorithm in one of programming language and the database files on the training set.
[4] After getting the proper result, the algorithm, developed any of the language and two data base files , using these two apply it on online reviews.

![Implementation Proposed Workflow Approach](image)

**EXAMPLE -**

<table>
<thead>
<tr>
<th>GOOD WORDS</th>
<th>ASSIGNED WEIGHTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Phrase Words</td>
<td></td>
</tr>
<tr>
<td>Famous</td>
<td>4</td>
</tr>
<tr>
<td>Cheap</td>
<td>4.5</td>
</tr>
<tr>
<td>Useful</td>
<td>4</td>
</tr>
<tr>
<td>Reasonable</td>
<td>3</td>
</tr>
<tr>
<td>Applications</td>
<td>3.5</td>
</tr>
<tr>
<td>Reliability</td>
<td>3</td>
</tr>
<tr>
<td>Very Good</td>
<td>3</td>
</tr>
<tr>
<td>Two Phrase Words</td>
<td></td>
</tr>
<tr>
<td>Long Run</td>
<td>3.5</td>
</tr>
<tr>
<td>Samsung Mobile</td>
<td>2.5</td>
</tr>
<tr>
<td>Middle Class</td>
<td>4.5</td>
</tr>
<tr>
<td>Resale Value</td>
<td>3</td>
</tr>
<tr>
<td>More number</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1 :- Positive words in one & Two Phrase

<table>
<thead>
<tr>
<th>BAD WORDS</th>
<th>ASSIGNED WEIGHTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Phrase words</td>
<td></td>
</tr>
<tr>
<td>Maintenance</td>
<td>-2</td>
</tr>
<tr>
<td>Questionable</td>
<td>-3</td>
</tr>
</tbody>
</table>

Two Phrase Words |  
| Short time     | -1               |
| Does not       | -3               |

Table 2:- Negative words for one & two phrase

5. Implementation Design And Result Evaluation

![Implementation Design](image)

Here I have show the front end implementation design view of my proposed work. Also this figure shows the its implementation design of different attribute of the mobile product. Also user's can write there own review on there ways.
In this screen shot it shows result of the given review which is written by the user’s. Here its generate the result according to a different attribute of the particular product.

6. Comparison With Other Method

Naïve Bayes

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>60.85%</td>
</tr>
<tr>
<td>2000</td>
<td>88.45%</td>
</tr>
<tr>
<td>6000</td>
<td>88.85%</td>
</tr>
<tr>
<td>14000</td>
<td>88.22%</td>
</tr>
</tbody>
</table>

Table 3: Different size of Training Data for Naïve Bayes

When the size of training data is smaller (nearly 800-1200), the result still has good performance. We can see that accuracy is 93.6206%. But when training set becomes a little larger (3000), the result is not good as smaller one. We can see that the result is improved up slightly. Compared to size, the improvement of accuracy is relatively small, and we can just improve accuracy to 3% (93.6206% to 96.2669).

Support Vector Machine

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>60.85%</td>
</tr>
<tr>
<td>2000</td>
<td>88.45%</td>
</tr>
<tr>
<td>6000</td>
<td>88.85%</td>
</tr>
<tr>
<td>14000</td>
<td>88.22%</td>
</tr>
</tbody>
</table>

Table 2: Different size of Training Data for SVM

When the training set is small (800-1200), the result of SVM model is much poor than others. When we use case-insensitive to create tf-idf vector, the accuracy can improve up to 20% (60.854% -> 72.6285%), which means that it is important to combine the information of uppercase and lowercase together to increase the concept for a specific term (ex: free, Free, FREE). The other reason is that if we see “free” and “Free” as the same term, then the data frequency of free will increase, so that we won’t throw away such important feature.

7. Conclusion And Future Work

Here I have proposed an opinion mining approach using machine learning and supervised learning, part of speech in which, it will present user friendly and easy approach, for finding the views of the customer, whether it is negative or positive or neutral for the product. Here algorithm using supervised learning like naive Bayesian, its give good result.

In the SVM algorithm the training set is small the result of SVM model is much poor than others. Also in the Association Rule Word Set of items two (at least) or more is generated from Association mining. So there is no option for considering a single word using association concept. Association mining largely reduces the
number of words to be considered for classifying texts, keeping only words having association between them.

Here I have found that naive bayes gives good performance and accurate result when training data set is smaller. So it is best suitable for my proposed work.

8. References

[1] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman and William Hersh - “Knowledge Management, Data Mining, And Text Mining in Medical Informatics” - Management Information Systems Department.