International Journal of Innovations in Engineering and Technology
http://dx.doi.org/10.21172/ijiet.82.050

Sentiment Analysis using Naïve Bayes Classifier

Pooja jain
M.tech student, Department of CSE
DCRUST,Murthal

Neetu verma
Assistant Professor, Department of CSE
DCRUST,Murthal

Abstract—In recent years, the remarkableexpansion of web technologies, lead to an massive quantity of user generated information in online systems.This large amount of information on web platforms make them viable for use as data sources, in applications based on opinion mining and sentiment analysis.Sentiment analysishas become a vital part in today's era. Post massiveexpansion of web technology, reviews existing on net are in surplus quantity. It wouldbe more helpful to an individual or organization if these opinions serve accurate sentiment of the whole review/document. This paper implements naïve Bayes algorithm to categorize the sentence in positive, negative and neutral precisely. So,weexecuted the proposed technique and we evaluated its performance, and suggested instructions of enhancement.

Keywords—Opinion mining; Sentiment analysis; Feature identification; Naïve Bayes classifier

I. INTRODUCTION

Generally individuals and companies are always interested in other's opinion like if someone wants to purchase a new product, then firstly, he/she tries to know the reviews i.e., what otherpeople think about the product and based on those reviews, he/she takes the decision.Before the Internet, people would seek opinions on products and services from sources such as friends, relatives, or consumer reports.Postenormous development of web technology, user seek their results on web.The development of internethas been an exponential increase in the amount of information in online systems. These very large volumes of information are very difficult to process by individuals, leading to information overload and affecting decision-making processes in organizations. So,Sentiment analysis has gained lot of importance and seen rapid growth of research in Natural Language Processing.Sentiment analysis is the operation of understanding the intent or emotion behind a given piece of text. It is used todetermine the polarities of the contents into positive, negative and neutral for the product, userreviews, user comments, and etc.This is enhanced than reading a large number of reviews. He can also relate the summaries of views of different products, instead of reading a large number of reviews.

The framework consists of the data collection process, pre-processing techniques, Sentence opinion, and evaluation. It is shown below inFigure 1: Sentiment analysis process

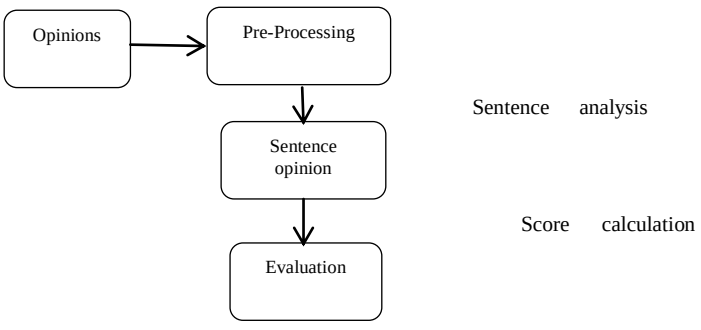


Figure 1:Sentiment analysis process

A. *Modules of OM*

OM model has different modules. The organizationof each module from unstructured review has been noticed by different research scholars. These constituentsintended to overcome the problem of queries arise through the mining process. Example – who write the opinion? Or what is the opinion? And the last is opinion about what??

So, based on these queries, modules are:

- "The source that has provided the opinion" is known as *Opinion holder*.
- "The attribute of entity about which opinion isconveyed" is known as *Target object/Feature*.
- And in last"Expression of opinion holder about the feature of the product" is known as *Opinion*.

B. *Kinds of OM*

- *Regular opinions*: It is frequentlymentioned as an opinion in the literature.it is further categorized in to two parts:

i. *Direct opinion:* A *direct opinion* refers to an opinion expressed directly on an entity or an entity aspect, e.g., "*The image quality is great*."

ii. *Indirect opinion:*An *indirect opinion* is an opinion that is conveyednot directly on an entity or phase of an entity based on its effects on some other entities.

- *Comparative opinions:*A comparative opinion expresses a relation of resemblances or difference between two or more entities. For example: coke tastes better than Pepsi.

II. SENTIMENT CLASSIFICATION LEVELS

In opinion mining, review is to be determined at three levels. These are:

- Sentence level classification.
- Document level classification.
- Aspect level classification.

A. *Sentence level classification*

This process involves two steps:

- Subjective classification in to one of two classes as objective and subjective
- Sentiment classification of subjective sentence in to three classes as positive, negative and neutral.

The job at this echelon goes to the sentences and decide whether each sentence depict a positive, negative, or neutral opinion. Neutral usually means no opinion. This levelis thoroughlyassociatedwith*subjectivity classification* , which differentiates sentences (called *objective sentences*) that express factual information from sentences (called *subjective sentences*) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions.

B. *Document level classification*

In this process, sentiments are extracted from entire document.The document level features are considered to classify the textual reviews on a single topic into positive, negative, and neutral. In general, the document features determines the overall sentiment polarity.

This works best when the document is written by single person or opinion holder or opinion is about single entity.

C. *Aspect level classification*

It is also known as Feature level classification. As an alternative of viewing at documents, paragraphs, sentences, clauses or phrases, aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a *sentiment* (positive or negative) and a *target* (of opinion). An opinion deprived of its

objective being acknowledged is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better.

III. NAÏVE BAYESAPPROACH

There are various methods used for opinion mining & sentiment analysis. But here we executed naïve Bayes classifier.

A. Introduction

The NAÏVE BAYES Classifier is well known machine learning method. It is probabilistic classifier given by Thomas Bayes. This technique assumes that the existence or nonexistence of any feature in the file is independent of the existence or nonappearance of any other feature. This basically helps in deciding the polarity of data in which opinions / reviews / arguments canbe classified as positive or negative which is facilitated by collection of positive or negative examples already fed.Naïve Bayes classifier believes a file as a bag of words and adopts that the probability of a word in the file is independent of its location in the file and the presence of other word. For a file f and class c:

$$p\,(c/f) = \frac{p(f/c)p(c)}{p(f)}$$

So, conditional probability of a sentiment is given as:

$$p(sentiment/sentence) = \frac{\mathrm{p}(sentence/sentiment)p(sentiment)}{p(sentence)}$$

a) Algorithm:

function TRAIN NAIVE BAYES(D, C) returns log P(c) and log P(w/c*)*
for each class c $\in C$ *# Calculate P(c) terms*
Ndoc = number of documents in D
Nc = number of documents from D in class c logprior[c]$\leftarrow log\ \frac{Nc}{Ndoc}$
V$\leftarrow$*vocabulary of D*
bigdoc[c]$\leftarrow$*append(d) for d* $\in$ *D with class c*
for each word w in V # Calculate P(w|c) terms count(w,c)$\leftarrow$*# of occurrences of w in bigdoc[c]*
loglikelihood[w,c]$\leftarrow log\frac{count(w,c)+1}{\sum w' in\ V\ (count(w',c)+1)}$

return logprior, loglikelihood, V

function TEST NAIVE BAYES(testdoc, logprior, loglikelihood, C, V) returns best c
for each class c $\in C$
sum[c]$\leftarrow$ *logprior[c]*
for each position i in testdoc
word$\leftarrow$*testdoc[i]*
if word $\in V$ *sum[c]*$\leftarrow$*sum[c]+ loglikelihood[word,c]*
return argmaxc sum[c]

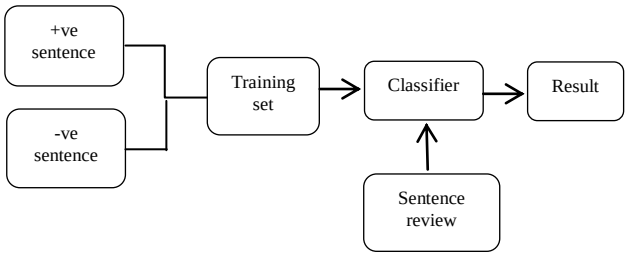
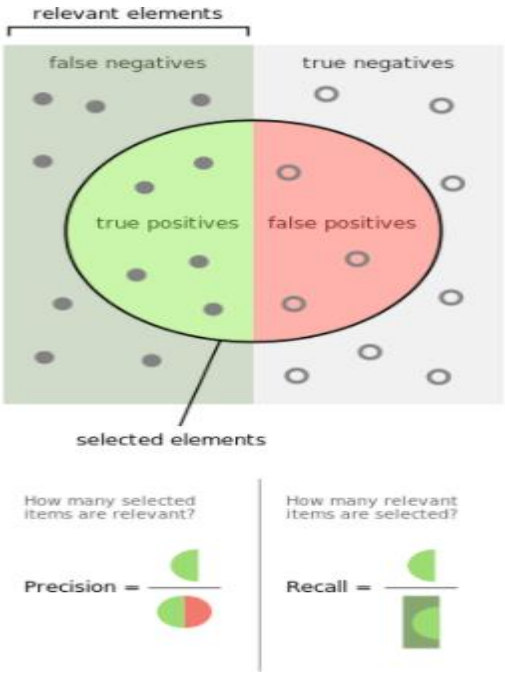


Figure 2: Algorithm of naive Bayes

*b) Evaluation:*To evaluate the algorithm following parameters are used:

- Accuracy:is the notch of closeness of magnitude of a quantity to that quantity's true value.
- Precision: **precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.
- Recall :**recall** (also known as sensitivity) is the fraction of relevant instances that have been retrieved over total relevant instances.
- Relevance: having any tendency to make the existence of any fact that is of consequence to the determinations of the action more probable or less probable than it would be without the evidence.

Following contingency pie chart is used to calculate the various measures.



c) Performance:

$$precision = \frac{tp}{tp + fp}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Recall = \frac{tp}{tp + fn}$$

d) Results:

Sentence	Sentiment	Probability of being positive	probability of being negative
Samsung phones has good camera quality	Positive	.703	.296
terrible movie	Negative	.274	.725
Phone is good but has low battery life.	Neutral	.673	.326

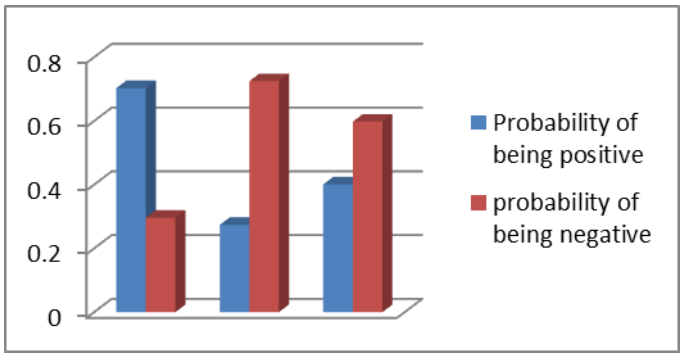


Figure 3:Results

B. Limitations:

Even though naïve Bayes technique is very simple and easy to implement. Still itclutches some concerns or limitations. These are:

- *Incomplete Training data:*In order to execute it, we need to calculate several conditional probabilities. Precisely, the class conditional probability, which defines the probability that an attribute suppose a specific value, given the consequence or reply class. In the standard naive Bayes instance of cricket data, there are no instances of "Play = No" when the trait "outlook" is "cloudy".So the class conditional probability would be zero and the entire construction breakdowns.
- *Continuous Variables*: When acharacteristic is continuous, calculating the probabilities by the traditional technique of frequency counts is impossible. In this case we would either need to transform the characteristic to a discrete variable or use probability density functions to calculate probability densities (not actual probabilities!).
- *Attribute Independence*: This is by farthe most important flaw and something which obliges a little bit of extra effort. In the calculation of consequenceprobabilities using the conventional Bayes theorem, the implicit assumption is that all the traits are mutuallyliberated. This allows us to multiply the class conditional probabilities in order to compute the outcome probability.

IV. Conclusion

The expression of opinions of consumers in specialized sites for estimation of products and services, and also on social networking platforms, has become one of the crucial ways of communication, due to remarkableexpansion of web environment in recent years. This paper presents a method of sentiment analysis, on the review made by users. Classification of reviews in both positive and negative classes is accomplished based on a naive Bayes algorithm. As training data we used a collection (pre-classified in positive and negative)

of sentences taken from the reviews. Our experiments results show that our method is very effective over existing method. In future work, we will improve our consequences and we will work on implicit features.

REFERENCES

[1] Bing Liu, 2012, Sentiment analysis and opinion mining, Morgan and Claypool publishers.
[2] B. Pang et al, 2002, Thumbs up : sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 79-86.
[3] P.D. Turney, 2002, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), 417–424.
[4] Riloff, E &Wiebe, J., 2003, Learning extraction patterns for subjective expressions, EMNLP˝03.
[5] Loren Terveen et al, 1997, PHOAKS: A system for sharing recommendations, Communications of the Association for Computing Machinery (CACM), 40(3):59–62.
[6] Minqing Hu and Bing Liu, 2004, Mining and summarizing customer reviews, Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining.
[7] Nasukawa, Tetsuya and Jeonghee Yi, 2003, Sentiment analysis: capturing favourability using natural language processing, Proceedings of the K-CAP03, 2nd International Conference on knowledge capture.
[8] Dave et al, 2003, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, In Proceedings of the 12th International Conference on World Wide Web, WWW 2003, 519-528.
[9] WiebeJanyce, 1990, Identifying subjective characters in narrative, Proceedings of the International Conference on Computational Linguistics (COLING-1990).
[10] Hearst M., 1992, Direction-based text interpretation as an information access refinement in Text-Based Intelligent Systems, P. Jacobs, Editor 1992, Lawrence Erlbaum Associates, 257-274.
[11] WiebeJanyce, 1994, Tracking point of view in narrative, Computational Linguistics, 233–287.
[12] Hatzivassiloglou et al, 1997, Predicting the semantic orientation of adjectives, Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997).
[13] Junichi Tatemura, 2000, Virtual reviewers for collaborative exploration of movie reviews, In Proceedings of Intelligent User Interfaces (IUI), 272–275.
[14] S. Morinaga et al, 2002, Mining product reputations on the web, SIGKDD˝02, Edmonton, Alberta, Canada.
[15] P.D. Turney and Michael L Littman,2003, Measuring Praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems, TOIS 2003, 21(4), 315-346.
[16] Esuli, A., &Sebastiani, F., 2005, Determining the semantic orientation of terms through gloss classification, In CIKM ˝05: Proceedings of the 14th ACM international conference on information and knowledge management, 617–624.
[17] Ion SMEUREANU, Cristian BUCUR, Applying Supervised Opinion Mining Techniques on Online User Reviews, InformaticaEconomică vol. 16, no. 2/2012.
[18] Nilesh M. Shelke, ShriniwasDeshpande, Vilas Thakre, Survey of Techniques for Opinion Mining, International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.