International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

Opinion Mining using Naïve Bayes Classifier

Vrinda
M.tech student, YMCAUST
Department of Computer Engineering
Faridabad, India

Dr. Komal Kumar Bhatia
Professor, YMCAUST
Department of Computer Engineering
Faridabad, India

***Abstract*— Opinion mining and Sentiment analysis have become a vital part in today's era. Both industries and consumers switch to online resources for feedback on products and amenities. Post enormous development of web technology, reviews existing on web are in surplus quantity. Thus, making decision problem harder. It would be more beneficiary to an individual or organization if these opinions serve precise sentiment of the whole review/document. This paper implements naïve Bayes algorithm to classify the sentence in positive, negative and neutral precisely. So, we implemented the proposed technique and we evaluated its performance, and suggested directions of improvement.**

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

I. INTRODUCTION

Opinion mining, also called Sentiment analysis, is the field of study that analyzes people's thoughts, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, events, topics, and their attributes. It can be defined as a method in which consumer reviews are mine for extraction of users. It is very challenging to mine opinions from reviews which are in natural language. Textual information can be divided into two main categories, facts and opinions. Facts are demonstrating objective statements about entities while opinions are signifying subjective statements that reflect people' sentiments. Opinion mining communicates with classifying opinion words, eg. brilliant, astonishing, optimistic, expensive, bad, and poor. Further opinion bearing words are classified into three orientations i.e., positive, negative or neutral. Opinions are so significant that whenever one requires making a decision, one wants to listen to others' views. If an individual wishes to buy a product, it is worthwhile to see a summary of sentiments of prevailing users so that he/she can make decision. 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. A fundamental phase in opinion mining & sentiment analysis application is feature extraction. The collective mining procedure is illustrated in Figure 1:Opinion mining process 1

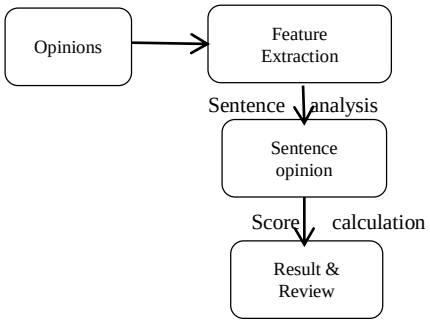

Figure 1: Opinion mining process

A. *Components of OM*

OM model has different components. The classification of each component from unstructured review has been addressed by different researchers. These components aimed to overcome the problem of questions 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 questions, components are:

- *Opinion holder:* The source who has provided the opinion.
- *Target object/Feature:* The attribute of entity about which opinion is conveyed.
- *Opinion:* Expression of opinion holder about the feature of the product.

B. *Types of OM*

- *Regular opinions*: It is often referred to basically 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 conveyed not 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.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

II. SENTIMENT CLASSIFICATION LEVELS

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

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

A. *Document level classification*

In this process, sentiments are extracted from entire document. And the opinion is categorized on overall sentiment of the opinion holder. The goal is to classify whole document as positive, negative or neutral.

Example "I purchased a Samsung Phone a couple days ago. It is a nice phone, even though a little large. The touch screen is cool. The video quality is clear too." Is the opinion classification positive or negative or neutral?

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

B. *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.

Subjective sentence signifies personal feelings, sights, emotions or belief. Just knowing that sentence is positive or negative is not enough as subjective sentence may contain multiple opinions.so, this is an intermediate step that helps filter out sentences with no opinion.

C. *Aspect/Feature level classification*

It executes finer-grained analysis. Instead of looking at constructs like sentence or document or clause, aspect level looks at the opinion straight. It is based on idea that reviews consists of sentiments as (positive/negative) and Target. Without target opinion is of limited use.so, recognizing the importance of target in an opinion helps us to comprehend sentiment analysis better.

III. NAÏVE BAYES CLASSIFICATION

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

A. *Introduction*

The NAÏVE BAYES Classifier is well known machine learning supervised method. It is probabilistic classifier given by Thomas Bayes. This classification technique assumes that the existence or nonexistence of any feature in the file is independent of the existence or nonappearance of any other feature. 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:

Step1: Initialize P(pos) ¬num - popozitii (positive)/ num_total_propozitii
Step2: Initialize P(neg) ¬num - popozitii (negative) / num_total_propozitii
Step3: Convert sentences into words

for each class of {pos, neg}:
for each word in {phrase}
P(word | class) < num_apartii (word | class) 1 | num_cuv (class) + num_total_cuvinte
P (class) ¬P (class) * P (word | class)
Returns max {P(pos), P(neg)}

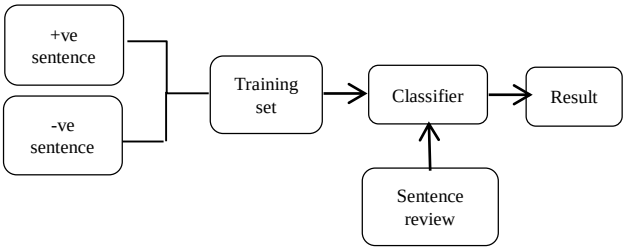

Figure 2: Algorithm of naive Bayes

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

- Accuracy
- Precision
- Recall
- Relevance

Following contingency table is used to calculate the various measures.

Table

	Relevant	Irrelevant
Detected opinions	True Positive (tp)	False Positive(fp)
Undetected opinions	False Negative(fn)	True Negative(tn)

c) Performance:

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

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

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

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

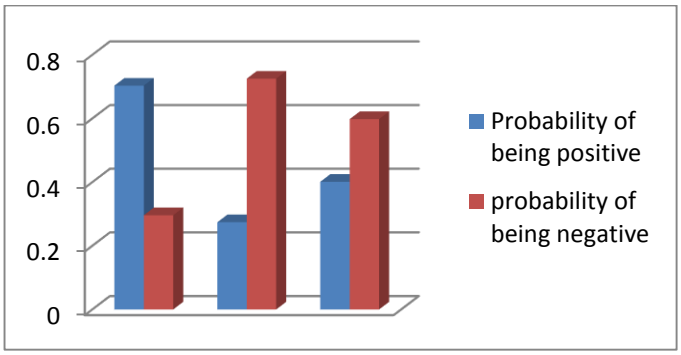

Figure 3:Results

B. Limitations:

Despite naïve Bayes technique is very simple and easy to implement. It still holds some issues or limitations. These are:

- *Incomplete Training data:* In order to implement it, we need to compute 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 a characteristic 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 far the most important flaw and something which obliges a little bit of extra effort. In the calculation of outcome probabilities using the classical Bayes theorem, the implicit assumption is that all the traits are mutually liberated. 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 remarkable expansion 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 results and we will work on implicit features.

V. 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.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

[17] Ion SMEUREANU, Cristian BUCUR, Applying Supervised Opinion Mining Techniques on Online User Reviews, Informatica Economică vol. 16, no. 2/2012.
[18] Nilesh M. Shelke, Shriniwas Deshpande, Vilas Thakre, Survey of Techniques for Opinion Mining, International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.