SAJET

South Asian Journal of Engineering and Technology Vol.2, No.15 (2016) 169–174

ISSN No: 2454-9614

Multiproduct Ranking through Naive Bayes Approach

D. Vidhya[a], G. Sivaselvan [a*], V.Vennila[a,b]

[a)] *Department Of Computer Science and Engineering, K.S.R. College of Engineering,, Tiruchencode, Namakkal, Tamilnadu, India.*
[b)] *Department Of Computer Science and Engineering, K.S.R. College of Engineering,, Tiruchencode, Namakkal, Tamilnadu, India.*
[c)] *Department Of Computer Science and Engineering K.S.R. College of Engineering,, Tiruchencode, Namakkal, Tamilnadu, India.*

*Corresponding Author: D. Vidhya

E-mail: dvidhya1993@gmail.com,

Received: 16/11/2015, Revised: 22/12/2015 and Accepted: 07/03/2016

Abstract

Online shopping is the recent trend in emerging world. It also provides the opportunity for customers to write reviews about products. For a popular product, the number of reviews can be in hundreds. Due to immense amount of customer's opinions, views and feedback available, it is very much significant to analyse and organize their views for better decision making. Opinion Mining or Sentiment Analysis is the mining of opinions and sentiments automatically from text, speech, and database sources through Natural Language Processing (NLP). Here all the customer reviews of a product are summarized. This summarization task is different from traditional text summarization because we are only interested in the specific aspects of the product that customers have opinions on and also whether the opinions are positive or negative. The frequent aspects are collected by Apriori algorithm. Naive Bayes classification and summarization technique classifies positive and negative opinions. Thus, ranking the given product based on the customer opinions will finalize the decision.

Keywords: Apriori Algorithm, Feature Extraction Techniques, Naive Bayes Classifiers, Opinion Mining, POS Tagging.
Reviewed by* **ICETSET'16 *organizing committee*

1. Introduction

Language is a powerful tool for communication. It is also a means to express emotion and sentiment. In current search engine people to search for other's opinions from the Internet before purchasing a product, when we are not aware with a specific product, we ask the trusted sources to recommend one. The web is a huge warehouse of structured and unstructured data. The analysis of this data to extract suppressed public opinion and sentiment is a challenging task. In data mining research field, machine learning techniques have been applied to automatically identify the information content in text.

2. Related Work

BakhtawarSeerat et al **[3]** proposed the work on how opinions are being extracted from online reviews and challenges of opinion mining. **[4]** Present an insight into task of opinion mining. Vijay B.Raut et al **[5]** have compared the methods and produced the summary of different approaches used for opinion mining and the results obtained. G.Vinodhini et al **[2]** presented an overview of different opinion mining techniques with approaches used. Pravesh Kumar Singh, MohdShahid Husain **[6]** investigated movie review mining using machine learning and semantic orientation.Grouping feature expressions, which are domain synonyms, is critical for effective opinion summary **[1].**

3. Opinion Mining

Opinion Mining or sentiment analysis can be defined as a sub-discipline of computational linguistics that emphases on extracting opinion of user. It is a Natural Language Processing (NLP) and Information Extraction (IE) task that goals is to obtain feelings of writer expressed in positive or negative comments by analysing a large number of documents.

The evaluation of opinion can be done in two ways:

• Direct opinion, gives positive or negative opinion directly about the object. For example, "The picture quality of this camera is poor" expresses a direct opinion.

• Comparison means to compare the object with other similar objects. For example, "The picture quality of camera-y is better than that of Camera-x." expresses a comparison.

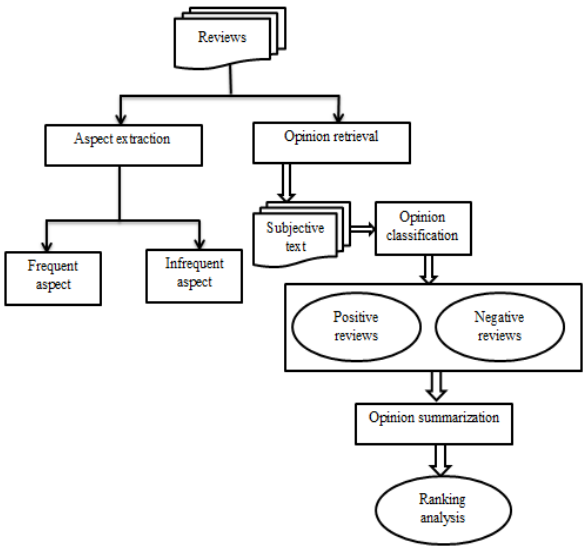

Fig.1 Workflow of Opinion Mining

3.1 Feature Extraction

Let us consider the n-gram features for feature extraction. An n-gram is a contiguous sequence of n items from a given sequence of text or speech. An n-gram could be any combination of letters **[7]** (syllables, letters, word, part-of-speech (POS), character, syntactic, and semantic n-grams). The n-grams typically are collected from a text or speech corpus and n-gram features captures sentiment cues in text. Fixed n-grams are exact sequences. Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. n-gram features can be classified into two categories:

1) Fixed n-grams are sequences occurring at either the character or token level.

2) Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. A plethora of fixed and variable n-grams have been used for opinion mining **[8].** Documents are often converted into vectors according to predefined features together with weighting mechanisms **[9].** Correlation is a commonly used method for feature selection **[10], [11].** The process of obtaining n–gram can be given as in the steps below,

1) Filtering – removing URL Links

2) Tokenization – Segmenting text by splitting it by spaces and punctuation marks, and forming bag of words

3) Removing Stop Words – Removing articles (“a”,”an”,”the”)

4) Constructing n-grams from consecutive words

3.1.1 Pre-Processing

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features.

Many feature subset selection algorithms, some can effectively eliminate irrelevant features form the reviews to eliminate the irrelevant word. Stop words are words which are filtered out before or after processing of natural language data. When building the index, most engines are programmed to remove certain words from any index entry. The list of words that are not to be added is called a stop list. Stop words are deemed irrelevant for searching purposes because they occur frequently in the language for which the indexing engine has been tuned. In order to save both space and time, these words are dropped at indexing time and then ignored at search time.

3.1.2 POS Tagging

The reviews are sent to the POS tagging module where POS tagger tags all the words of the sentences to their appropriate Part-of-Speech tag. POS tagging is an important phase of opinion mining, it is necessary to determine the features and opinion words from the reviews. POS tagging can be done manually or with the help of POS tagger. Manual POS tagging of the reviews take lots of time. Here, POS tagger is used to tag all the words of reviews.

3.1.3 Aspect Identifier and Frequent Aspect Process

Frequent features are the “hot” features that people comment most about the given product. However, there

are some features that only a small number of people talked about. These features can also be interesting to some potential customers and the manufacturer of the product. A priori algorithm is used in finding frequent item sets.

All the features are extracted from the reviews and stored in a dataset then its corresponding opinion words are extracted from these reviews. An object is an entity which can be a product, service, person, event, organization, or topic. It is associated with a set of components or attributes, called aspects of the object. Each component may have its own set of aspects. An opinion is simply a positive or negative view, attitude, or emotion about an object or an aspect of the object from a person or an organization. Given a collection of opinion texts on an object, the aspect extraction problem is to produce the aspects of the object from these documents.

The next is that opinion words are usually associated with aspects under certain syntactic relations. We can derive a set of aspects in terms of syntactic relations. Similarly, syntactic clues can help extract new aspects from the extracted aspects, and new opinion words from the extracted aspects. This propagation process continues until no more opinion words or aspects can be extracted.

3.2 Opinion Classification

3.2.1 Extracting opinion words and seed list preparation

Initially some of the common opinion words along with their polarity are stored in the seed list. All the opinion words are extracted from the tagged output. The extracted opinion words matched with the words stored in seed list .If the word is not found in the seed list then the synonyms are determined with the help of WordNet. Each synonym is matched with words in the seed list, if any synonym matched then extracted opinion word is stored with the same polarity in the seed list. If none of the synonym is matched then the antonym is determined form the Word Net and the same process is repeated, if any antonym matched then extract opinion word is stored with the opposite polarity in the seed list. In this way the seed list keep on increasing. It grows every time whenever the synonyms or antonyms words are found in Word Net matches with seed list.

3.2.2 Predicting the orientation of opinion sentences

The next and step of the process is predicting the orientation of an opinion sentence, i.e., positive or negative. In general, we use the dominant orientation of the opinion words in the sentence to determine the orientation of the sentence. That is, if positive/negative opinion prevails, the opinion sentence is regarded as a positive/negative one. In the case where there is the same number of positive and negative opinion words in the sentence, we predict the orientation using the average orientation of effective opinions or the orientation of the previous opinion sentence. We focus on the double propagation method which is based on the following observations. It is easy to identify the set of opinion words such as “good” and “bad,” etc.

3.2.3 Naive Bayes Algorithm

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution.

For example, suppose the training data contain a continuous attribute, x.

1) We first segment the data by the class.
2) Then compute the mean and variance of x in each class.
3) Let μ_c be the mean of the values in x associated with class c, and let σ_c^2 be the variance of the values in x associated with class c.
4) Then,the probability distribution of some value given a class, $p(x = v|c)$, can be computed by plugging v into the equation for a Normal distribution parameterized by μ_c and σ_c^2.
5) That is,

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Another common technique for handling continuous values is to use binning to discretize the feature values, to obtain a new set of Bernoulli-distributed features; some literature in fact suggests that this is necessary to apply naive Bayes, but it is not, and the discretization may throw away discriminative information.

3.3 Multiproduct Ranking

This approach, which includes Bayesian inference to rank the products based on the opinion mining. In existing system the opinion mining is used to find the aspect based classification and summarization. By using this method we just analysis the single product. We are going to propose a system for multi-product ranking using opinion mining. This system get the input as multi product reviews datasets in same time and extract the aspects and opinion and classify the opinion using Bayesian classifier. The aspect extraction, opinion extraction and opinion classifier are handled for different products in same time based on the reviews. This system automatically ranks the given products by using the option status.

4. Conclusion

In conclusion, this work has successfully extended an existing aspect-based mining approach in order to apply it to the tourism domain, particularly, to opinions available on the Web in the manner of products reviews. As a result of the new and more complex NLP-based rules that we developed for both subjective and sentiment classification, our extension is able to perform better than Liu's model, improving both Accuracy and Recall for the mentioned tasks. The effectiveness of these rules shows that the features that we detected on tourism products, such as sentences including multiple mentions of the product or the presence of a high number of sentences containing no opinions, are an accurate characterization of the domain and that they should be considered in future work on the field for a good performance.

The objective is to produce a feature-based summary of a large number of customer reviews of a product sold online. This problem will become increasingly important as more people are buying and expressing their opinions on the Web. Our experimental results indicate that the proposed techniques are effective in performing their tasks. Our method also handles implicit features represented by feature indicators. These make the proposed technique more complete. Experimental results show that the proposed technique performs markedly better than the state-of-the-art existing methods.

5. Future Work

There are several challenges in analysing the sentiment of the web user reviews. First, a word that is considered to be positive in one situation may be considered negative in another situation. Take the word "long" for instance. If a customer said a laptop's battery life was long, that would be a positive opinion. If the customer said that the laptop's start-up time was long, however, that would be a negative opinion **[7].** These differences mean that an opinion system trained to gather opinions on one type of product or product feature may not perform very well on another.

References

[1] Zhai Z, Liu B, Xu H, and Jia P, Grouping Product Features Using Semi-supervised Learning with Soft-Constraints, in Proceedings of COLING. 2010.

[2] G.Vinodhini et al, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol 2, Issue 6, June 2012.

[3] BakhtawarSeerat, FarouqueAzam, "Opinion Mining: Issues and Challenges (A Survey)," International Journal of Computer Applications, Vol49 No 9 July 2012, Pg No 42-51.

[4] N.Mishra and C.K.Jha, "An insight into task of opinion mining," Second International Joint Conference on Advances in Signal Processing and Information Technology – SPIT 2012.

[5] Vijay .B.Raut et al, "Survey on Opinion Mining and Summarization of User Reviews on Web", International Journal of Computer Science and Information Technologies (IJCSIT),Vol 5(2), 2014.

[6] Pravesh Kumar Singh, MohdShahidHusain , "Methodological Study Of Opinion Mining And Sentiment Analysis Techniques," International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014.

[7] Ahmed Abbasi, Stephen France, Zhu Zhang and Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 3, pp. 447- 462, 2011.

[8] Michael Wiegand and Alexandra Balahur, "A Survey on the Role of Negation in Sentiment Analysis", Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, 2010.

[9] Yuming Lin, Jingwei Zhang, Xiaoling Wang and Aoying Zhou, "Sentiment Classification via Integrating Multiple Feature Presentations", WWW 2012 – Poster Presentation, pp. 569-570, 2012.

[10] M. Hall and L.A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach", Proceedings of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems, pp. 855- 858, 1997.

[11] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research**,** Vol. 3, pp. 1289-1305, 2004.