International Journal of Engineering

Journal Homepage: www.ije.ir

RESEARCH NOTE

Mining Interesting Aspects of a Product using Aspect-based Opinion Mining from Product Reviews

K. Srividya*[a], K. Mariyababu[b], A. Mary Sowjanya[b]

[a] *Department of CSE, Gmrit ,Rajam, India*
[b] *Department of CS&SE, Andhra University College of Engineering (A), Visakhapatnam, AP, India*

PAPER INFO

Paper history:
Received 26 May 2017
Received in revised form 17 August 2017
Accepted 08 September 2017

Keywords:
Sentiment Analysis
Opinion Mining
Aspect Term
Aspect Based Analysis
Customer Review

ABSTRACT

As the internet and its applications are growing, E-commerce has become one of its rapid applications. Customers of E-commerce were provided with the opportunity to express their opinion about the product on the web as a text in the form of reviews. In the previous studies, mere founding sentiment from reviews was not helpful to get the exact opinion of the review. In this paper, we have used Aspect-Based Opinion Mining to get more interesting aspects of a product's sentiment from unlabelled textual data. First, noun phrases algorithm was used to get all the aspect term of a review sentence. Secondly, the sentiment algorithm was applied on the result of the noun-phrase algorithm and also applied on adjectives and on adverbs. Finally, using relative importance algorithm important aspects were presented to the user. Our proposed methodology has achieved 77.03% of accuracy compared to previews studies. The proposed methodology can be applied for any product reviews in the form of text without any label, and it does not require any training dataset.

***doi**: 10.5829/ije.2017.30.11b.11*

1. INTRODUCTION

As the internet and its technology is growing, people got the freedom of expressing their views, interests and opinions about the things they see around or use regularly in the form of reviews and feedback. Now, a day's lots of people are using the internet and doing online shopping and eventually they are going to look for good things. Today's service providers or product providers are more interested in the reviews of their customers because they contain the opinion of the customer and/or, his/her interest about that product or service. Service providers are faced challenging issues in finding behaviour or interest of their customer.

Since people had forum discussions, blogs, Twitter, comments, and postings in social network sites, it requires special attention of data analysts in the company to maintain successive growth in their business. They need to analyse people's interests about a particular entity (ex., product).

*Corresponding Author's Email: srividya.kotagiri@gmail.com (K. Srividya)

Phrase level sentiment analysis or aspect-based sentiment analysis is the best solution to mine people's interests from online reviews. Aspects are the attributes of the service or product and we have named service or product as an entity.

Aspect-based sentiment analysis is context dependent solution since reviewers used a different type of text in a different type of media they are using. Phrase level sentiment analysis/ opinion mining works by finding all the aspect terms and gets an opinion about those mined aspects.

For example, consider a sentence "I impressed for dell customer service but it is getting motherboard problems". In this sentence, aspect terms are 'customer service' and 'motherboard'. Opinion or sentiment expressed towards those aspects is 'impressed' and 'problems'. By using our school English knowledge we can say that customer is satisfied with dell's customer service but unsatisfied with their motherboard performance.

Using phrase-level sentiment analysis we can solve the problems of data analysts to make important

Please cite this article as: K. Srividya, K. Mariyababu, A. Mary Sowjanya, Mining Interesting Aspects of a Product using Aspect-based Opinion Mining from Product Reviews, International Journal of Engineering (IJE), TRANSACTIONS B: Applications Vol. 30, No. 11, (November 2017) 1707-1713

decisions about their products or services. To get the sentiment of the text we have to make a machine to learn and it could accomplish in two ways, one is supervised learning and unsupervised learning. For supervised learning class labels of each text is known before doing any classification. For unsupervised learning class label is unknown and it will be found later.

Classification predicts the class labels of categorical data. Class labels are predetermined and it builds the classifier model. For some data we train the classifier then we perform testing for new data. This classifier tries to find out the relationship among attributes and classifies based on the maximum relationship in terms of probability.

Supervised text classification algorithms for their famous are Naïve Bayes (Gaussian NB, Multinomial, and Bernoulli NB), Support Vector Machines and Maximum-Entropy classifier. The accuracy of classification is different for these algorithms and way of classification is also different.

Naïve Bayes classification algorithms are the more scalable algorithm to classify documents based on the frequency of words. Naïve Bayes classifiers consider more number of features for training the classifier. Naïve Bayes algorithm follows Bayes theorem by assuming independence of its features.

Ample of the study happened on Naïve Bayes classifiers at 50’s and it was called with many names at 60’s for text classification. It needed a number of linear in parameters and features in learning phase for learning cases. Support Vector Machines (SVM) was another classifier in the competition in this field. Naïve Bayes uses maximum likelihood training. Naive Bayes was more competitive than SVM in text classification problems.

Naïve Bayes is fast and more scalable in model building and in scoring the text. It can be used to classify both Binary and multiclass problems of classification. It works using conditional probability. It calculates the probability by the ratio of combination (pair) wise occurrence with the solo (single) occurrence of features.

The paper is organized as follows: In section 2, background and related work is reviewed. Section 3 specifies how to get aspects which are more in polarity based on their relative importance. Section 4 works by getting aspect terms from each sentence. In sections 5 and 6, experimental results are provided with evaluation measures. Paper is finally concluded in section 7.

2. BACKGROUND KNOWLEDGE

The sentiment is the subjective term which refers opinion of the customer which covers context and target of a particular entity. In this research, we mainly discuss all the previous studies about sentiment analysis. As the first one, traditional text classification is also known as sentiment analysis for thumbs up/down reviews.

The document-level sentiment analysis was used to study the polarity of the whole document as the subtask of document classification based on polarity. The result of this study was just POSITIVE/ NEGATIVE about the document. Initial work for this purpose had Turney [1] and Pang et al. [2] to work greatly to get the polarity of product reviews.

Work of Turney was documented level. Classification for document polarity can behave on multi-way scale, and this work was carried out by Pang and Lee [3] and Snyder and Barzilay [4]. However, Pang and Lee [3] improved this fundamental task of classifying positive and negative for predicting star labels of 3 and 4 scale, meanwhile Snyder and Barzilay [4] applied an extensive analytics of restaurant reviews, predicting on various aspects of hotel such as service and food etc. (on a five-star scale).

However, number of statistical classification strategies, the class neutral was ignored by assuming that neutral class lies near the boundary line, but many of the researchers say that, like any sentiment or polarity problem, 3 categories must be found or identified. In spite of all these, most of the particular classifiers such as the Max Entropy [5] and the SVMs [6] can able to get more accuracy and improve its classification performance.

In one of the two ways, we can deal with the neutral class labels. Either, identification of neutral language by the algorithm, and removing all the neural words and collecting words like positive and negative, or the algorithm builds a classification in a three-way [7]. The second approach sometimes needs the estimation of the probability distribution of all the class labels (e.g. Naive Bayes classifiers as implemented by Python's NLTK kit).

The way of using neutral data highly depends on given data: getting the polarity of negative and positive is quite easy if the language is labeled as neutral, negative and positive by extracting all the negative languages. If the data is most neutral in contrast with positive and negative polarities then it is highly harder to find the difference between positive and negative polarities of the language.

The sentence-level sentiment analysis was used to get sentiment as a sentence wise. It was trying to find a number of positive and negative words in a sentence. If positive words is less than negative words then the sentence is negative else positive.

The phrase level/aspect based opinion mining refers determining the opinion or sentiment expressed on different features or aspects of entities, e.g., of an iPod, a cell phone, or an online shopping mall [8]. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, the service for a restaurant, or the picture quality of a camera. The use of aspect-based opinion analysis is the possibility to get the poorness about objects of interest.

Various aspect can give various opinion responses, for example, a restaurant can have a good location, but normal food [9]. This issue (problem) include many sub-issues (problems), e.g., identifying related entities (target), getting their attributes/features/aspects, and finding whether a sentiment given on each attribute/feature/aspect is positive, negative or neutral [10]. The dynamic (automatic) identification of attributes/features can be applied with methods syntactic or modeling of the topic [11, 12]. Extensive details about this level of opinion analysis can be studied in Liu's work [13]. According to Chintalapudi and Prasad [14], sentiment analysis can be on communities by identifying small real-world networks manually, whereas in large scale networks it is an extremely difficult problem. Naïve Bayes is fast and more scalable in model building and in scoring the text. It can be used to classify both binary and multiclass problems of classification [15].

3. PROBLEM STATEMENT

Users are unable to observe more interestingness of the reviewers by using document and sentence opinion mining. In aspect-based opinion mining, users get polarity aspects and even not able to get the aspects preferred by the customers. The goal of the task is to get aspects which are more in polarity based on their relative importance. The proposed system identities polarity of the aspect and then gets most related aspects in descending order of their relative importance.

4. PROPOSED SYSTEM

The proposed system works by getting aspect terms from each sentence. Finding the aspect terms is a little-complicated process, so here we consider nouns as the aspect terms in the noun phrase. Noun phrases are extracted based on the chunk grammar using chunk parser. Chunk grammar used here is as follows:

ChunkGrammar

(r"NP:{<NN.?>+<RB><JJ>|" "<NN><RB><JJ>|"
"<JJ><NNS><RB>|"
"<RB>+<JJ><NN><RB>|"
"<RB><NN.?>+<VB.?>*<NN.?>+<RB>|"
"<IN><RB><JJ>|"
"<NN><JJ><RB>|"
"<RB><VB.?>+<JJ><NN>|"
"<RB><NN.?>+<VB.?>+<RB>|"
"<JJ><NN><RB>*|"
"<RB>|"
"<JJ><NN><RB><JJ>}")

It is not the standard but based on the dataset and its domain will get changed. Representation above is called regular expression representation.

Algorithm to find Noun Phases

Input: POS-Tagged words
Output: a Tree with noun phrases
Method: Extract noun phrases for dataset D of sentence S_i according to the chunk grammer rule
Do: FOR EACH SENTENCE S_i DO

Figure 1 explains how the proposed system process the given product reviews and finally find out and displays the most interested aspects.

The following section gives the detailed flow of the proposed word. The proposed system uses customer reviews and goes through a data preprocessing stage which deals with stop word removal, stemming, and lemmatization. Here, for each sentence get noun and other words as separate. After that, for each aspect count the probabilities of other words. According to the sum, aspect will be categorized.

A. Stop Word Removal

All the frequent used words in the English language which are not useful in finding any opinion in data mining are called stop words.

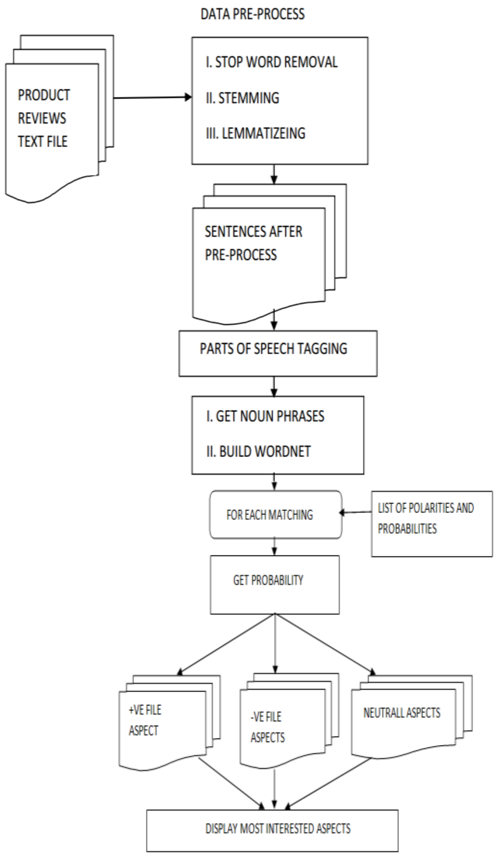

Figure 1. Working of Proposed Architecture

Stop words are language dependent and do not carry any information about text with them. They may be pronouns, prepositions, conjunctions. To delete each unwanted word in each review sentence, this stops word removal is used. Words are “the, at, are, was”, etc. Reviews are stored in a text file which is given as input to stop word removal. Stop words are collected and stored in a text file. Stop word is removed by checking against stop words list.

B. Stemming

The way of getting root word from given word is called stemming. A stemming algorithm cuts the words "worry", "worried", and "worrying" to the root word, "worry". It includes many algorithms like n-gram analysis, affix stemmers and lemmatization algorithms. Porter stemmer algorithm is used to form root word for given input reviews and store it in a text file.

C. Lemmatization

Lemmatization is the same as the stemming but it works well for some words than steamer; also get root word for words not stemmed by stemmer.

Algorithm for Data-Preprocess

Input: a text file containing all the reviews
Output: sentences free from stop words, and stemmed, lemmatized words.
Method: Tokenize the given data set D into sentences $S_1,S_2,S_3,\ldots S_i$ again tokenize the sentence into words $W_1,W_2,W_3,\ldots W_j$
For each word W in a sentence, S look for the presence of it in stop word. If you found it, skip that word else include it and go for another word.
Do: FOR EACH SENTENCE in S_i DO

D. POS Tagging

The syntactic or morphological behavior of a word is called POS-Tagging based on its linguistic category. Most of the POS-Tagging categories in English grammar are a noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. POS-Tagging is the job of attaching each word in a sentence with its appropriate POS-tag or just tagging.

POS-Tagging is treated as the most salient part in opinion mining. It is important to identify attributes or features, opinion words from review sentence. POS-Tagging can be done either static (manually) or with the help of POS-Tagger tool. Manual POS-Tagging of the reviews is time-consuming. POS-Tagger tags all the words from reviews. NLTK’s POS tag was used to tag each word in reviews. Everyone sentence in customer reviews is tagged and stored in a text file. The following table shows an algorithm for POS-Tagging.

Algorithm for POS-Tagging

Input: POS-Tagged words
Output: a Tree with noun phrases
Method: Extract noun phrases for dataset D of Sentence S_i according to the chunk grammer rule
Do: FOR EACH SENTENCE S_i DO

E. Noun Phrases

Noun phrases are formed for English language and as we assumed that aspect can be found in noun phrases and formed noun grammar to accomplish this task. Algorithm for noun phases shows the algorithm to find noun phrases.

F. Build Word-net

Word-net is the synonyms for a word. This is very much useful in the proposed system. Because many words which users were written may not be directly present in List of polarity and Probabilities.

G. Aspect Classification Algorithm(Get Probabilities)

Algorithm for Aspect Classification

Input: noun list, other list, word probabilities
Output: Aspect classification as positive, negative and neutral
Method:
get noun list NL, other list OL
For each item in OL build wordnet ON
For each item in ON look in word probabilities
If True get probability and polarity
Repeat
Sum up all probabilities is noun/ noun list
    If sum >=0.7 it is positive
    If sum <=0.4 it is negative
    Else it is neutral
Repeat

5. EXPERIMENTAL SETUP

The dataset we used was provided with NLTK package product reviews Corpora in Python. The experiment was done with iPod product. Data was given in text format. All the positive, negative and neutral sentences were mixed. For any dataset like above, the proposed system works well. All the data is unlabelled and mix of different sentences. For word probabilities, we have used vender_lexicon which was provided with NLTK sentiment corpora. This lexicon contains all the words with their polarity and probability.

6.RESULTS AND DISSCUSSION

The proposed system is dealing with the aspects of given product reviews. Initially, the proposed algorithm extracts the noun phrases and find outs the probability of that noun (it is polarity of the noun), based on the probability nouns are separated. Finally, nouns are statistically measured, and plots the most 10 interested aspects. The interestingness of an aspect is nothing but number of times it is repeating in respective polarity.

Figure 2 shows the most interesting positive aspects from the product reviews. In the diagram relative importance is nothing but the number of times s particular aspect is repeated. Here, we have taken the 10 most interesting aspects. By doing this, we can say that users are referring particular aspect more number of times in their reviews.

Figure 3 shows the most interested negative aspects from the product reviews. To get the most interesting negative aspects we find out the negative noun probability and have taken 10 most repeating nouns to plot them.

Figure 4 shows the most interesting neutral aspects from the product reviews. These aspects are separated based on the noun probability.

A. Extracting interesting aspects from product reviews using aspect based opinion mining.

These are the statistics for our data set:

Total sentences = 300

Total noun phrases = 603

Total not nouns = 711

Total relevant nouns = 335

Total nouns = 350

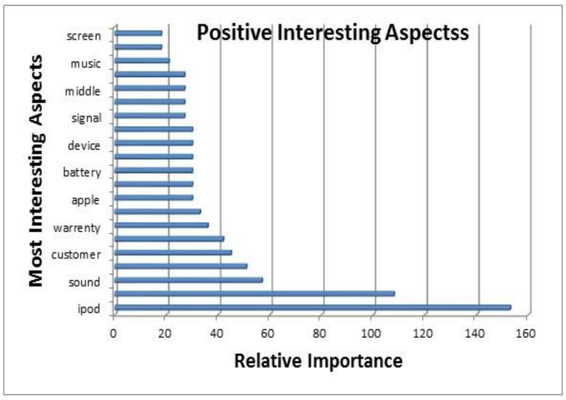

Figure 2. Most Positive Aspects

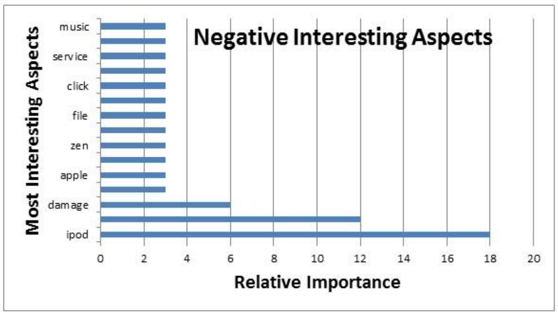

Figure 3. Negative Aspects

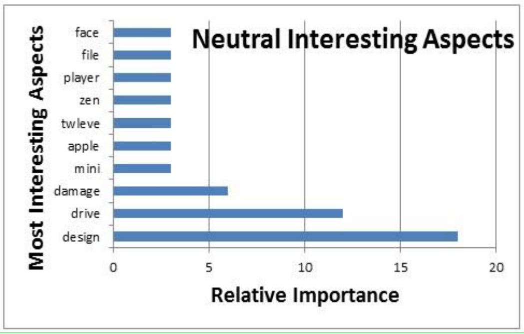

Figure 4. Neutral Aspects

B. Evaluation measures for aspect (Attributes) extraction:

Precision = (total noun phrase + total nouns selected) / total nouns

= (603 + 350)/ 1061

= 89.82%

where, 1061 (711 + 350)

Recall = Relevant nouns selected / Total nouns selected

= 335/350

= 95.71%

F-Measure = (17193.3444) / (89.82 + 95.71)

=(17193.3444)/185.53

=92.67%

C. Evaluation measures for most relevant aspects:

Total positive nouns = 139

Total negative nouns = 29

Total neutral nouns = 68

Total conflict nouns = 9

Total = 245

Precision = (139+29+68)/245

= 96.32%

Recall = (139+29)/(139+29+9)

=94.91%

F-Measure = (2(96.32 * 94.91))/(96.32 + 94.91)

= 18283.46/191.23

= 95.57%

D. Evaluation measures for aspect extraction and most interesting aspects (see Figure 5)

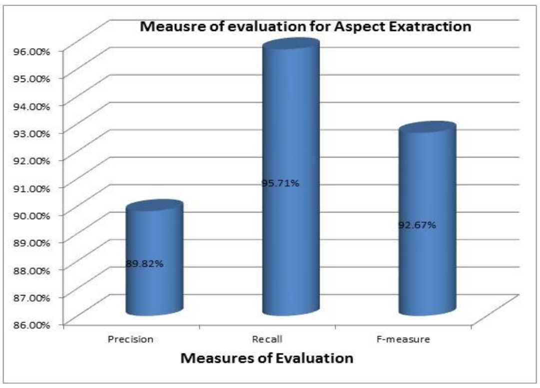

Figure 5. Measures for Aspect Extraction

Evaluation measure for getting most interesting aspects
Figure 6 measures for interesting aspect extraction
Comparison of the previous works
Total reviews taken = 100
Total sentences taken = 400
Positive sentence = 231
Negative sentence = 108
Total opinion sentences = 339

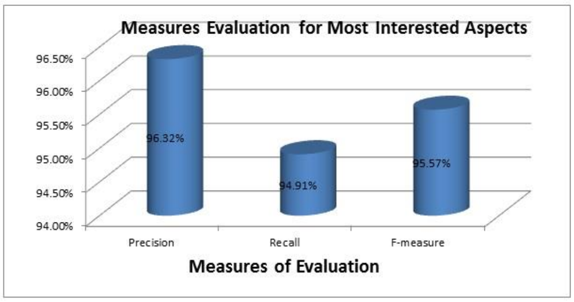

Figure 6. Measure for Getting Most Interesting Aspects

7. CONCLUSION AND FUTURE WORK

The proposed system extracts the most related aspects from the product reviews and suggests the user about those aspects. The probabilities which were already computed using naïve bays algorithm are used to find the polarity of the each aspect. Finally, based on aspect frequency most related aspects were founded. In the future, it is proposed to predict the aspect from different contexts.

8. REFERENCES

1. Turney, P.D., "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews", in Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics., (2002), 417-424.
2. Pang, B., Lee, L. and Vaithyanathan, S., "Thumbs up?: Sentiment classification using machine learning techniques", in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics., (2002), 79-86.
3. Pang, B. and Lee, L., "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", in Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics., (2005), 115-124.
4. Snyder, B. and Barzilay, R., "Multiple aspect ranking using the good grief algorithm", in HLT-NAACL., (2007), 300-307.
5. Vryniotis, V., "The importance of neutral class in sentiment analysis", Machine Learning Blog & Software Development News, (2013).
6. Koppel, M. and Schler, J., "The importance of neutral examples for learning sentiment", ***Computational Intelligence***, Vol. 22, No. 2, (2006), 100-109.
7. Gonçalves, P., Araújo, M., Ribeiro, F., Benevenuto, F. and Gonçalves, M., "A benchmark comparison of state-of-the-practice sentiment analysis methods", ***arXiv preprint arXiv:1512.01818***, (2015).
8. Hu, M. and Liu, B., "Mining and summarizing customer reviews", in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. (2004), 168-177.
9. Cataldi, M., Ballatore, A., Tiddi, I. and Aufaure, M.-A., "Good location, terrible food: Detecting feature sentiment in user-generated reviews", ***Social Network Analysis and Mining***, Vol. 3, No. 4, (2013), 1149-1163.
10. Liu, B., Hu, M. and Cheng, J., "Opinion observer: Analyzing and comparing opinions on the web", in Proceedings of the 14th international conference on World Wide Web, ACM., (2005), 342-351.
11. Zhai, Z., Liu, B., Xu, H. and Jia, P., "Constrained lda for grouping product features in opinion mining", ***Advances in Knowledge Discovery and Data Mining***, (2011), 448-459.
12. Titov, I. and McDonald, R., "Modeling online reviews with multi-grain topic models", in Proceedings of the 17th international conference on World Wide Web, ACM., (2008), 111-120.
13. Liu, B., "Sentiment analysis and subjectivity", ***Handbook of Natural Language Processing***, Vol. 2, (2010), 627-666.
14. Chintalapudi, S.R. and Prasad, M.K., "Mining overlapping communities in real-world networks based on extended modularity gain", ***International Journal of Engineering-Transactions A: Basics***, Vol. 30, No. 4, (2017), 486-492.
15. Vennila, G., Shalini, N.S. and Manikandan, M., "Navie bayes intrusion classification system for voice over internet protocol network using honeypot", ***International Journal of Engineering TRANSACTIONS A: Basics***, Vol. 28, No. 1, (2015), 44-51.

Mining Interesting Aspects of a Product using Aspect-based Opinion Mining from Product Reviews

RESEARCH NOTE

K. Srividya[a], K. Mariyababu[b], A. Mary Sowjanya[b]

[a] *Department of CSE, Gmrit ,Rajam, India*
[b] *Department of CS&SE, Andhra University College of Engineering (A), Visakhapatnam, AP, India*

P A P E R I N F O

Paper history:
Received 26 May 2017
Received in revised form 17 August 2017
Accepted 08 September 2017

Keywords:
Sentiment Analysis
Opinion Mining
Aspect Term
Aspect Based Analysis
Customer Review

چکیده

همانطور که اینترنت و برنامه های کاربردی آن در حال رشد هستند، تجارت الکترونیک یکی از کاربردهای سریع آن شده است. مشتریان تجارت الکترونیک با این فرصت برای ابراز نظرات خود در مورد محصول در وب به عنوان یک متن به شکل بررسی ایجاد شده اند. در مطالعات قبلی، احساسات پایه ای از بررسی ها به تنهایی برای نظر دقیق در مورد بررسی مفید نبودند. در این مقاله، از مفهوم نظرات مبتنی بر ابعاد استفاده کرده ایم تا جالب تر از احساسات محصول از داده های متنی بدون برچسب باشد. اولا، الگوریتم عبارات اسم برای بدست آوردن تمام جنبه های یک جمله بازبینی استفاده شد. دوم، الگوریتم احساسات در نتیجه الگوریتم عبارات اسم مورد استفاده قرار گرفت و همچنین بر روی صفت ها و قیدها اعمال شد. در نهایت، با استفاده از الگوریتم اهمیت نسبی، جنبه های مهم به کاربر ارائه شد. روش پیشنهادی ما نسبت به مطالعات پیشین به ٪۷۷/۰۳ دقت رسید. روش پیشنهادی می تواند برای بررسی هر محصول به صورت متن بدون برچسب استفاده شود و نیازی به مجموعه داده های آموزشی ندارد.

***doi**: 10.5829/ije.2017.30.11b.11*