ICIC Express Letters
Volume **14**, Number **11**, November **2020**

ICIC International ©2020 ISSN 1881-803X
pp. 1033–1040

LANGUAGE MODEL COMBINED WITH WORD2VEC FOR PRODUCT'S ASPECT BASED EXTRACTION

Tu Nguyen Thi Ngoc[1], Ha Nguyen Thi Thu[1,*] and Viet Anh Nguyen[2]

[1]Information Technology Faculty
Electric Power University
234 Hoang Quoc Viet, Hanoi 100000, Vietnam
*Corresponding author: hantt@epu.edu.vn

[2]Department of Data Science and Applications
Vietnam Academy of Science and Technology
18 Hoang Quoc Viet, Hanoi 100000, Vietnam

Received January 2020; accepted April 2020

Abstract. *Opinion mining has provided significant information to the business and the other fields. Based on user perspective, it is possible for enterprises to improve their products to suit their customers' requirements or be able to innovate in line with consumer trends and it can suggest to customer another equivalent product. In this paper, we present a method for identifying the product aspects from customers' comments based on semi-supervised learning technic. The main idea uses Word2Vec for representing semantic and determining the similar words. Firstly, we train with core terms and Word2Vec model, and after that we use support measure of words with each aspect for expecting what aspect a sentence can belong to. The experiments are carried out using a dataset for hotel review collected from TripAsvisor.com and results show that our method is really effective with approximate accuracy 80%.*

Keywords: Sentiment, Opinion, Aspect extraction, Word2Vec, Semi-supervised, Supporting

1. **Introduction.** With the explosion of e-commerce and multimedia technology during the last two decades, comments on the Internet have become more popular. Comments of users express their opinion on products and services, such as hotels, digital products, coffee, and beer. These comments make large data (called big data) and high speed (social big data). For processing them, we collected reviews from websites and analyzed for multi-purpose. Thus, they attracted the attention of researchers in the exploitation of larger social data to describe, identify and predict customers' behaviour in a number of areas, especially in transaction trade on the Internet.

Opinion mining has three main tasks: 1) defining (extracting) aspects; 2) classification; 3) estimating of the aspect weights. Aspect definition is the task of extracting aspects of an entity that is mentioned in the comments. Opinion classification is the task of determining the type of class this comment belongs to (positive, negative or neutral). In some cases, opinion classification can become a critical task in the ranking (i.e., quantitative) aspects. Weighted aspect estimation is the task of determining the difference importance of each aspect product in the users' view.

Recently, some studies focus on aspect extraction. Most of them are based on frequent nouns and noun phrases [1,2] for identifying and extracting single aspect but ignore aspects that have low frequent nouns or noun phrases. Other studies use extracting method based on rule [3-6], and they try to overcome the limit of frequency method. Several supervised-learning methods [7,8] are used for extracting aspects. These supervised learning methods

DOI: 10.24507/icicel.14.11.1033

cannot build a general model for all area products. Recently, some studies that applied topic modeling [9-11] can detect latent aspect in comments, but it is limited when they need a large data and need to adjust data in training process. All of these methods did not mention to the context of words.

In this paper, we use a lexical analysis approach to extract aspects, and then we use W2V model to calculate similarity context of words. After that, we calculate their weights based on the supporting measure. Weighted words will be aggregated and predicted in next tasks.

This paper is structured as follows: Section 2 presents related work, Section 3 presents the proposed methodology of building core terms and aspects extraction, Section 4 shows experiment and evaluation of the proposed method and finally conclusion section concludes the paper and gives some future research directions.

2. **Related Work.** Aspect extraction is an important task in the opinion mining field. Some proposed methods determined aspect based on frequency [1,2], in which nouns or noun phrases with high frequency are candidates for extracting and calculating. In [1], nouns and noun phrases are assigned by the Part-of-Speech (POS), and after that they used the data mining algorithm to extract the nouns with high frequency. The bias is added from experiment to re-select nouns again. It seems quite effective but in some cases, low frequency nouns and noun phrases are missed although, they are really significant for determining an aspect. In [2], authors used baseline statistics of words in English and probability-based heuristics to more accurately identify features of product categories. The Feature Extractor begins by counting the total number of times n_x that each noun x appears in reviews of products in category c. Then system also computes the total number of noun occurrences ($N = \sum_x n_x$) in the categories' reviews. Next, the algorithm calculates the probability that lemma x would occur n_x times in random English text containing a series of N noun occurrences.

To overcome this problem, a number of study built sets of rules for determining aspects, and these methods are called rule-based methods [3-6]. In [6], they tried to remove noun phrases that they cannot be an aspect of entity. They are calculated by Point-wise Mutual Information (PMI) between phrases and meronym that has related to the entity. For example, meronym for camera layer is: "of camera", "camera has", "camera comes with", etc. In [5], authors used Term Frequency – Inverse Document Frequency (TF-IDF) combined with a filter based on pattern (pattern-based filter) for removing non-aspect terms. In Long et al.'s study [4], they performed to extract aspects based on frequency and information distance. Firstly, core terms are detected by the frequency-based method. After that, related words are extracted based on information distance [3]. For example, the price aspect includes symbols as "$" or "dollars". Limitation of this method creates non-aspects that they matched related patterns.

To process these problems, a number of studies used machine learning model to train from suggestion data like Hidden Markov Model (HMM) [7] and Conditional Random Field (CRF) [8]. In [7], a system called Opinion Miner was developed based on aspect extraction and related opinion by using Lexicon-HMM (L-HMM). Li et al. [8] integrated two parameters of CRF, Skip-CRF and Tree-CRF, for extracting aspects and mining opinion. Not the same as original CRF, Skip-CRF and Tree-CRF can permit CRF mine syntax features. Although machine learning-based methods are more effective than rule-based method and frequency-based method, it still needs a set of entities for assigning and training.

The unsupervised learning models are concerned in some studies. These researches are primarily based on two main basic models: Probabilistic Latent Semantic Analysis (PLSA) [9] and Latent Dirichlet Allocation (LDA) [9]. In [9], Blei et al. proposed to use the basic LDA model for document classification. In particular, LDA is used as a dimensionality reduction method, as it reduces any document to a vector of real-valued features, i.e., the posterior Dirichlet parameter associated with each document. The parameters of an LDA model are learned using all the documents, without reference to their true class label. Brody and Elhadad [10] proposed LDA model in the sentence level to determine the local topic of each sentence as an aspect. Then identify the specific semantics in each aspect through the adjectives. Moghaddam and Ester [11] introduced the Factorized LDA (FLDA) on sentence-level to extract aspects. Wang et al. [14] proposed two semi-supervised learning models FL-LDA (Fine-grained Labelled LDA) and UFL-LDA (Unified Fine-grained Labelled-LDA) for extracting aspect from comments. In [12], authors developed a hidden-aspect extraction model, which used the Support Vector Machine (SVM) and topic modeling based on LDA. LDA is used to extract aspects, in which several aspects are assigned labels to different topics before applying topic modeling. Unsupervised learning uses datasets without labelled; however, they still need large data and need to adjust for achieving reasonable results.

3. Methodology.

3.1. Problem definition.

Definition 3.1. Set of comments. *Set of comments is a set of reviews about a type of product* $D = \{d_1, d_2, \ldots, d_N\}$.

Definition 3.2. Aspect. *Aspect is a set of values that describe about a type of product* $A = \{A_1, A_2, \ldots, A_K\}$.

Definition 3.3. Word2Vec. *Word2Vec is an unsupervised learning model, which is trained from a large corpus. Word2Vec has 2 models: skip-gram and Cbow [15,16].*

3.2. Modeling. Formally, several studies calculated the similarity between two sentences based on Manhattan, Euclidean measure. In our paper, we used a supporting measure of words combined with Word2Vec for calculating sentence score and determining an aspect of a sentence as Figure 1. In this methodology, reviews were collected by a set of reviews. We use Word2Vec to train, and after that, we extracted a set of word vectors. With this approach, we can reduce the cost for building sets of words and the cost of time for processing.

We have several sets after training:
$S = \{w_1, w_2, \ldots, w_N\}$ is collected words w_i extracted from sentence S;
$A = \{A_1, A_2, \ldots, A_K\}$ is a set of aspects (A_j is also called topic);
$A_{corej} = \{w_1, w_2, \ldots, w_t\}$ is core term set for each aspect A_j.

3.3. Aspect extraction based on supporting measurement. In Figure 1, we describe 2 phases of the method as follows.

✠ *Training phase*

- ***Step 1 (Data):*** Set of customers' comments about hotels is separated into sentences, pre-processing sentences and save notional words.
- ***Step 2 (Training with Word2Vec):*** In this step, we use the Word2Vec model for training, and python is a tool to extract words that have similarity semantic with each other.
- ***Step 3 (Building core terms for each aspect):*** We perform to classify and assign aspects to words (we called aspect core terms).
- ***Step 4 (Topic word extraction):*** In this step, word vector in Step 2 will extract in to topic word set if it does not belong to the core term of aspect.
- ***Step 5 (Calculating supp(word → aspect)):*** The supporting measure of word w_i is calculated in aspect A_j like Equation (1). Equation (1) is improved Euclidean

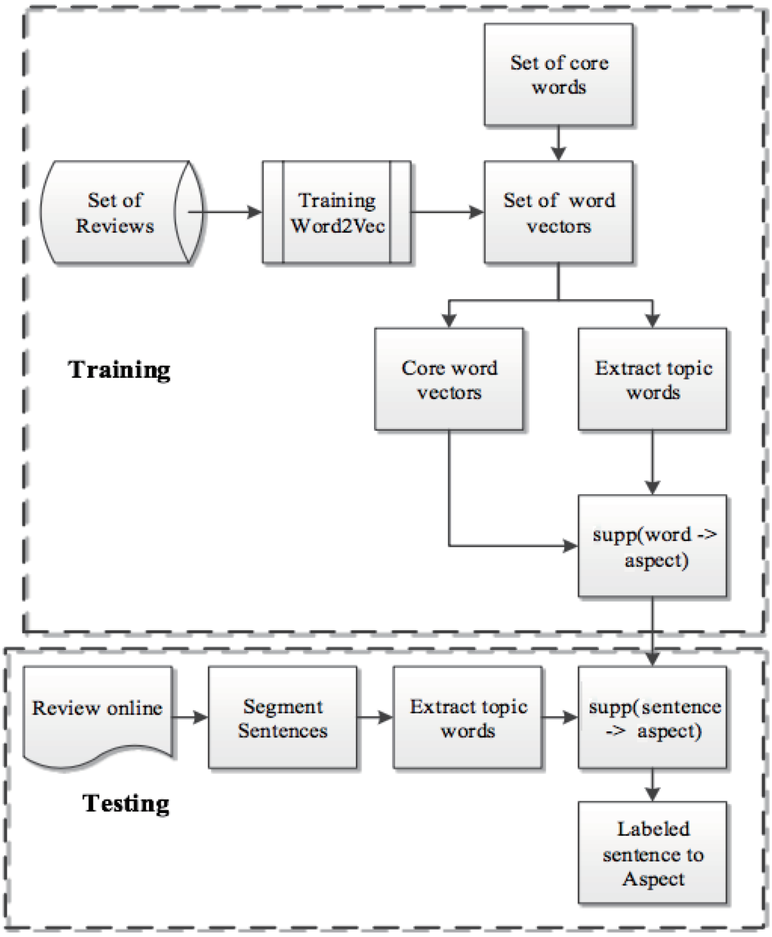

FIGURE 1. Estimate sentences' aspect based on language model combined with Word2Vec

measure.

$$supp(w_i \to A_j) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\sum_{k=1}^{K}\left(x_{kw_i} - x_{kcore_{tA_j}}\right)^2} \tag{1}$$

in which:

$supp(w_i \to A_j)$: supporting of w_i to aspect A_j,
n: number of core terms for aspect A_j,
K: number of dimensionals of a word,
x_{kw_i}: aspect of dimension kth of word w_i,
$x_{kcore_{tA_j}}$: value of dimension kth of core term tth that belongs to aspect A_j.

Figure 2 illustrates a word set that is built based on supporting measure where A_j ($j = 1, 2, \ldots, 7$) is the aspects and symbols $a, b, c, \ldots, w$ are the words. The arcs of A_j and words express the supporting measure of words for each aspect A_j, in which supporting $(x_i \to A_j)$ and $(x_i \to A_h)$ are different.

✠ *Testing phase*

Step 1 (Sentence segmentation)

Step 2 (Word extraction): extracting notional words in sentence.

Step 3 (supp(sentence → aspect)): Matching extracted word with trained word for determining supporting word to aspect. Based on the supporting measure of words to aspect, we can calculate supporting of a sentence to aspect by following formula

$$supp(S \to A_j) = \frac{1}{P}\sum_{i=1}^{P} supp(w_i \to A_j) \tag{2}$$

in which:

$supp(S \rightarrow A_j)$: supporting of sentence S to aspect A_j,
$supp(w_i \rightarrow A_j)$: supporting of word w_i to aspect A_j;
P: number of words in sentence S.

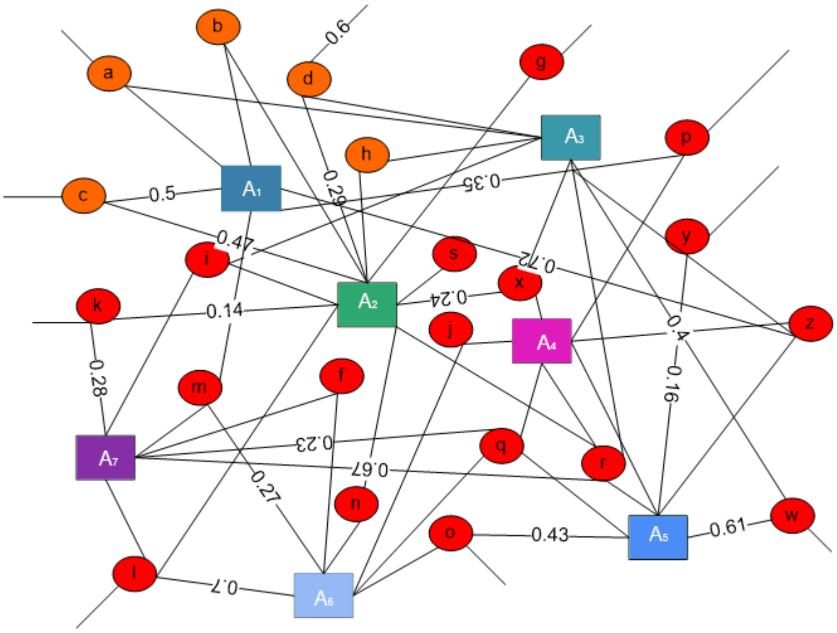

FIGURE 2. Supporting of words to aspects

```
Aspect extraction based on supporting measure algorithm
Input:
– $D$: A comment
– $A$: $A = \{A_1, A_2, \ldots, A_K\}$: Set of aspects
– $Core$: $Core = \{core_1, core_2, \ldots, core_N\}$: Set of core terms corresponding with aspect
$A_j$ $(j = 1 \div N)$
– $S$: $S = \{s_1, s_2, \ldots, s_M\}$: Set of sentences of comment $D$.
Output:
  $S_j$: set of sentences that is assigned with label $A_j$ $(j = 1, \ldots, N)$
Initialization:
  $S_j = \Phi$ // $j = 1, \ldots, N$
  $S \leftarrow split(D)$
For each $s_m \in S$ do
  {float max = 0;
    For each $A_j \in A$ do
      {For $i = 1$ to length $(s_m)$ do
        {Calculate supporting of $w_i$ to each aspect $A_j$ by Formula (1)
        }
       Calculate supporting of sentence $s_m$ to each aspect $A_j$ by Formula (2)
      }
    If (max <= $supp(s_m \rightarrow A_j)$) then
      {max = $supp(s_m \rightarrow A_j)$;
       $s_m \leftarrow label(A_j)$; $S_j \leftarrow s_m$;
      }
  }
```

4. Experiment.

4.1. Training set.

For experiment, we used a training set from TripAdvisor about hotel. After collecting data, we perform data cleansing (remove the noisy symbol, correct mistake).

In the hotel data have seven aspects. Table 1 illustrates some core terms corresponding with 7 aspects.

Table 1. Aspects of hotel

Aspect	**Core term**
Value	Value, price, worth
Room	Room, rooms
Location	Location
Cleanliness	Dirty, smelled, clean
Check in/front desk	Staff
Service	Service, breakfast, food
Business service	Internet, wifi

4.2. Evaluating.

There are many ways/measurements to evaluate the accuracy and effectiveness of classification. In this paper, we use two measurements for evaluating method, including: recall and precision. In addition, we built a tool that is developed from .Net framework and SQL Server. Table 2 shows the results of our proposed method.

Table 2. Results of aspect extraction

Aspect	**Precision**	**Recall**
Value	0.77	0.750
Room	0.78	0.750
Location	0.82	0.790
Cleanliness	0.76	0.72
Check in/front desk	0.80	0.80
Service	0.73	0.68
Business service	**0.85**	**0.83**

We compared our method with two other baseline methods LDA [9] and Long et al. [4] on the same corpus. Precision is used for evaluating. The results in Table 3 and Figure 3 show that our method outperforms LDA with a large margin. Our method outperforms Long et al.'s in value, location, cleanliness, check in, and business service aspects. However, Long et al.'s method outperforms us in detecting the service aspect.

Table 3. Comparison with LDA and Long et al.

Method	Value	Room	Location	Cleanliness	Check in/ front desk	Service	Business service
LDA	0.65	0.47	0.56	0.60	0.65	0.59	0.60
Long et al.	0.76	0.78	0.75	0.75	0.74	0.75	0.75
Our method	0.77	0.78	0.76	0.76	0.80	0.73	0.85

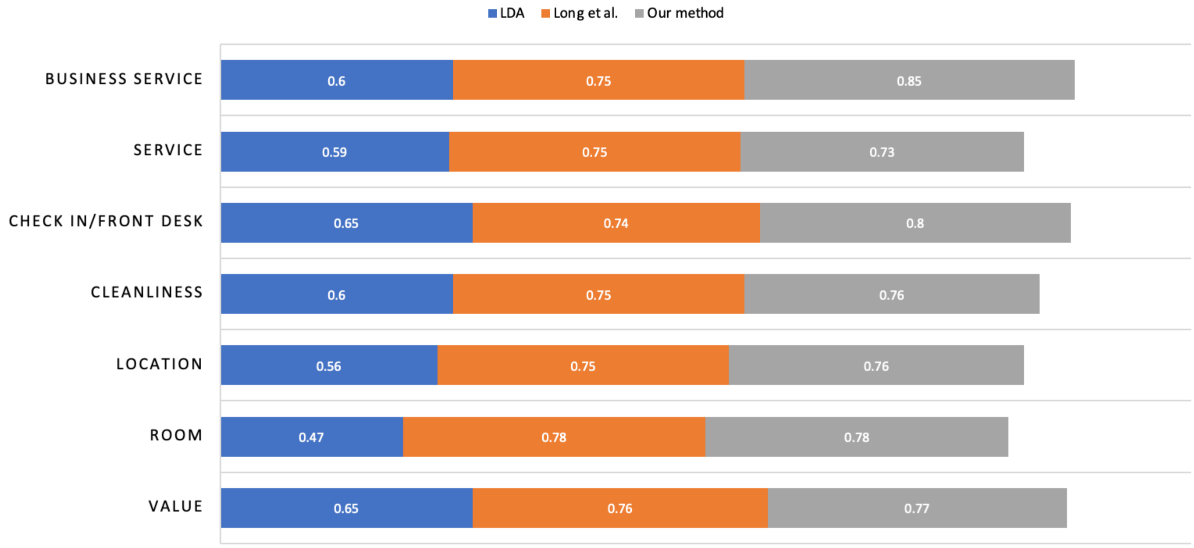

Figure 3. Comparison with LDA and Long et al.'s methods

5. **Conclusion.** Aspect extraction is the core task for sentiment analysis, so that there are many studies focusing on this problem in recent years. In this paper, we have proposed a model for product aspect extraction. In it, a dictionary was developed from Word2Vec, and we calculated weight of the core term by the supporting measure. The experimental results have shown that, it works very well on real world datasets in comparison with other state-of-the-art methods, and it can be applied for some different fields.

In the future, we will adapt training model with deep level, and improve parameter for the better results.

REFERENCES

[1] M. Hu and B. Liu, Mining and summarizing customer reviews, *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004.

[2] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng and C. Jin, Red Opal: Product-feature scoring from reviews, *Proc. of the 8th ACM Conference on Electronic Commerce*, San Diego, CA, USA, 2007.

[3] R. L. Cilibrasi and P. M. B. Vitanyi, The google similarity distance, *IEEE Trans. Knowledge and Data Engineering*, vol.19, no.3, pp.370-383, 2007.

[4] C. Long, J. Zhang and X. Zhu, A review selection approach for accurate feature rating estimation, *Proc. of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010.

[5] S. Moghaddam and M. Ester, Opinion digger: An unsupervised opinion miner from unstructured product reviews, *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, ON, Canada, 2010.

[6] A. M. Popescu and O. Etzioni, Extracting product features and opinions from reviews, *Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 2005.

[7] W. Jin, H. H. Ho and R. K. Srihari, OpinionMiner: A novel machine learning system for web opinion mining and extraction, *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009.

[8] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang and H. Yu, Structure aware review mining and summarization, *Proc. of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010.

[9] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[10] S. Brody and N. Elhadad, An unsupervised aspect-sentiment model for online reviews, *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, CA, USA, 2010.

[11] S. Moghaddam and M. Ester, The FLDA model for aspect-based opinion mining: Addressing the cold start problem, *Proc. of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, 2013.

[12] H. Xu, F. Zhang and W. Wang, Implicit feature identification in Chinese reviews using explicit topic mining model, *Knowledge-Based Systems*, 2014.

[13] T. Hofmann, Probabilistic latent semantic indexing, *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information*, Berkeley, CA, USA, 1999.

[14] T. Wang, Y. Cai, H.-F. Leung, R. Y. K. Lau, Q. Li and H. Min, Product aspect extraction supervised with online domain knowledge, *Knowledge-Based Systems*, vol.71, pp.86-100, 2014.

[15] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, *arXiv: 1301.3781*, 2013.

[16] *https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2vec/*, Accessed in June, 2019.

[17] L. B. Khanh, H. N. T. Thu and T. D. Thanh, Deep level Markov chain model for semantic document retrieval, *Scalable Information Systems*, vol.5, no.19, DOI: 10.4108/eai.19-6-2018.155443, 2018.