MINISTRY OF EDUCATION AND TRAINING
VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

NGUYEN THI NGOC TU

RESEARCH ON IMPROVING SOME ASPECT-LEVEL SENTIMENT ANALYSIS BASED ON MACHINE LEARNING

Major: Information System

Major code: 9 48 01 04

SUMMARY OF INFORMATION SYSTEM THESIS

HÀ NỘI - 2023

The thesis has been completed at Graduate University of Science and Technology- Vietnam Academy of Science and Technology

Supervisor: **Prof. Dr. Nguyen Viet Anh**

Reviewer 1: …
Reviewer 2: …
Reviewer 3: ….

The thesis shall be defended in front of the Thesis Committee at Vietnam Academy Of Science And Technology - Graduate University Of Science And Technology, at ….… hour……, date…… month….… year 2023.

This thesis could be found at:

- The National Library of Vietnam
- The Library of Graduate University of Science and Technology

INTRODUCTION

1. The urgency of the thesis

In recent years, a lot of people often express their opinions about things such as products and services on social networks and e-commerce websites. These opinions or reviews often play a significant role in improving the quality of products and services. Exploiting these reviews can understand the views and psychology of consumers, thereby helping to build business strategies such as marketing campaigns, priority products, and monitoring. It can also be made to learn consumer behavior, and market patterns, and predict society's consumption trends.

Because of the importance of opinion mining, over the past two decades, researchers, academics, organizations, and businesses have been interested in researching this field. At present, aspect-based sentiment analysis is attracting much attention from the research community and application developers. Aspect-based sentiment analysis, a systematic synthesis of views about entities and their attributes can be generated. This turns unstructured text into structured data and is usable for all types of qualitative and quantitative analysis.

The two main tasks in aspect-based sentiment analysis are *aspect extraction* and *aspect sentiment classification*. The growing interest in opinion mining and aspect-level sentiment analysis is partly due to its potential applications. However, there are still many challenges including:

- For aspect extraction task:
  The first challenge is the lack of labeled training data for this task. Second, many sentences in the review lack explicit aspect expressions (nouns) making it difficult to identify aspects. In addition, there are many ways of alluding to aspects (hidden features) appearing, which can make the aspect extraction task even more complicated. Third, when a word appears, its context should be considered. For many, their meanings depend on the context in which they are used. For example, the word “apple” appears in two sentences: “Apple is a tasty fruit” and “Apple has just launched a new product” are understood in two different meanings. Fourth, there are some important aspects that have a low frequency that is easily overlooked. How to detect such aspects is also a challenge of the aspect extraction task.
- For the aspect sentiment classification task:
  First, the multi-class sentiment classification task is more challenging than the two-class classification. The presence of multiple classes makes it more difficult for a classifier to determine the margins between

different classes. Second, the closeness between sentiment classes or between the classes that have the same sentiment polarity is almost similar, and they are easy to misclassify each other. Third, a word can have different meanings based on the context and domain in which it is used. The meaning of the same word can be different for each context. For example, the word "long time" when talking about the battery life of the phone has a positive opinion, but in the context of the CPU processing speed, it has a negative opinion. Finally, the presence of negation can reverse the emotional polarity of a text. However, it is not easy to handle this by reversing polarity because negative words can be found in a sentence without affecting the emotion expressed in the text.

From the survey and evaluation of the research, the author realizes that there is a need for a full study on all tasks of aspect-based sentiment analysis to provide useful background information for practical applications. It is necessary to find an effective approach to overcome the challenges in the field of research, improving the performance of the aspect-based sentiment analysis system. The author of the thesis chooses the topic "*Research and develop some machine learning algorithms in economic forecasting*".

2. The objectives of the thesis

The objective of the thesis is to propose a system to perform three tasks of the online product evaluation aspect-level sentiment analysis problem. From there, the Ph.D. student proposes some semi-supervised machine learning algorithms to extract aspects and sentiments of customer express in each aspect and proposes some supervised machine learning algorithms to solve the aspect sentiment classification.

3. The main contents of the thesis

The thesis studies the background knowledge in sentiment analysis and the problem of aspect-based sentiment analysis.

The thesis studies traditional and modern machine learning methods and proposes 02 semi-supervised algorithms to extract aspects and sentiments from online product reviews. The first algorithm is based on conditional probability combined with bootstrapping algorithm, the second algorithm is based on WordtoVector representation. The Ph.D. student also proposes Naïve Bayes, Support Vector Machine, OR gate Bayesian network, and Dempster-Shafer combination theory for facet emotion classification task.

CHAPTER 1. OVERVIEW OF SENTIMENT ANALYSIS AND ASPECT-BASED SENTIMENT ANALYSIS

1.1 Overview of sentiment analysis

1.1.1. Some basic knowledge

Definition 1.7 Opinion: An *opinion* is a quintuple (e_i, a_{ij}, s_{ijkl}, h_k, t_l). where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k.

- The opinion defined is called a regular opinion.

1.1.2. Sentiment Analysis Tasks

Task 1 (entity extraction and categorization)

Task 2 (aspect extraction and categorization): Extract all aspect expressions of the entities, and categorize these aspect expressions into clusters. Each aspect expression cluster of entity e_i represents a unique aspect a_{ij}.

Task 3 (opinion holder extraction and categorization)

Task 4 (time extraction and standardization)

Task 5 (aspect sentiment classification): Determine whether an opinion on an aspect a_{ij} is positive, negative, or neutral, or assign a numeric sentiment rating to the aspect.

Task 6 (opinion quintuple generation): Produce all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) expressed in document *d* based on the results of the above tasks.

1.1.3. Sentiment analysis levels

Document-level sentiment analysis: At this level, the process aims to classify whether a whole document expresses a negative or positive sentiment or opinion.

Sentence-level sentiment analysis: At this level, the focus is on the sentence. The main goal is to determine whether the sentence expresses a positive, negative, or neutral opinion.

Aspect-level sentiment analysis: This level performs fined-grained analysis because it aims to find sentiments with respect to the specific aspects of entities.

1.1.4. Feature problems in natural language processing and opinion analysis

1.2 Aspect-based sentiment analysis

1.2.1. Process of Aspect-based sentiment analysis

1.2.2. Aspect-based sentiment analysis tasks

There are three main tasks in aspect-based opinion mining. Those tasks are (1) aspect extraction; (2) aspect sentiment classification; (3) summarizing opinion based on the aspect. There is also a subtask for determining the aspect weight (the importance of an aspect).

Aspect extraction: identify all aspect terms contained in each sentence of the review or the entire review text.

Aspect sentiment classification: for an aspect, determine the polarity of each aspect term or the aspect as a whole.

Estimating aspect weight: This task is to estimate the non-negative weights of the aspect that a user places on the aspect of the review.

1.2.3. **Aspect extraction methods**

Explicit aspect extraction*:* can be classified into three types according to the learning approach: unsupervised, semi-supervised, and supervised.

- Unsupervised techniques include *Frequency or statistical methods*, *Heuristic- or Rule-based methods*, and *Pointwise mutual information (PMI)*.
- The semi-supervised approaches include *Bootstrapping*, *Dependency parser*, and *Lexicon-based methods*.
- Supervised techniques include *Lexicalized HMM-based*, *Conditional random field (CRF)*, *Recurrent neural network* (RNN), and *Convolutional neural network* (CNN).

Implicit aspects extraction*:* can be classified into unsupervised, supervised, and hybrid approaches.

- Unsupervised methods are further divided into co-occurrence-based, topic modeling-based, clustering-based, and other methods.
- Supervised methods include classification-based, rule-based, and sequence tagging-based methods.
- Hybrid approach applied a combination of methods for the task of implicit aspect detection.

1.2.4. Methods of classifying aspect sentiment

Most literature usually divides sentiment analysis approaches into three categories: Machine Learning approaches, Lexicon-Based approaches, and Hybrid approaches.

Machine Learning Approach: these include *supervised learning*, *unsupervised learning*, *semi-supervised learning*, *reinforcement learning*, and *deep learning*.

- There are four types of supervised classification approaches which are linear, probabilistic, rule-based, and decision tree.
- Unsupervised approaches in the field of sentiment analysis generally use clustering: *hierarchical* and *partition* clustering.
- Recent works on *semi-supervised learning-based* sentiment analysis can be classified into five categories generative, co-training, self-training, graph-based, and multi-view learning.
- Reinforcement learning (RL) is a machine learning method where an agent is rewarded in the next time step based on the evaluation of its previous action.
- The term ‘‘deep learning’’ refers to neural networks with multiple layers of perceptron inspired by our brain. DL includes many neural network models such as CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), and DBN (Deep Belief Networks).

Lexicon-based approach: also called knowledge-based approach. The three major techniques for creating and annotating sentiment lexicons are the manual approach, the dictionary-based approach, and the corpus-based approach.

- The manual approach requires human intervention to annotate the lexicon.
- Building sentiment lexicons based on the corpus method, a dictionary is learned from data with a statistical and semantic approach.
- The hybrid approach combines both lexicon and machine learning approaches.

1.3 Some machine learning knowledge is used in the thesis for aspect-based sentiment analysis

1.3.1. Bootstrap technique

1.3.2. Word to vector

1.3.3. SVM classification

1.3.4.Naïve Bayes classification

1.3.5. Disjunctive interaction Noisy OR-gate

1.4 Evaluation metrics of sentiment analysis

CHAPTER 2 ASPECT-LEVEL SENTIMENT ANALYSIS

2.1 Introduction of problem

Users often refer to different aspects, that is the attributes or components of the product. For each aspect, users often give their opinion by expressing a positive or negative attitude about that aspect.

How to understand the content of the review and the issues that users mention? Aspect-based sentiment analysis solves that problem in detail aspects of the product that users mentioned in their reviews. This level indicates which aspects of the user's review were mentioned by the user, the customer's sentiment/opinion on each of those aspects, and finally the level of interest of each customer on each aspect.

The aspect-based sentiment analysis problem consists of three subproblems: (1) The aspect extraction problem; (2) The aspect sentiment classification problem ; (3) The aspect weighting problem.

The aspect extraction problem is to identify all the aspects that appear in the user's reviews. There are several challenges in this task as follows: Some aspects are explicitly mentioned and some are not. It is necessary to extract both explicit and implicit aspects. How to minimize noise while still being able to identify rare and important aspects is also solved.

It is assumed that the universal set of all possible aspects for each product is readily available together with aspect words called core terms (terms that describe aspects). This assumption is practical because the number of important aspects is often small and can be easily obtained by domain experts. The aspect extraction task then becomes how to correctly assign existing aspects to sentences in the review. The main challenge here is that in many reviews, sentences do not contain enough core terms or even do not have any core term at all, and thus may be assigned with wrong aspects. This problem is solved by repeatedly updating and enlarging the set of core terms to the set of aspect words by using the conditional probability technique combined with the bootstrap technique. Then, a supervised approach called the Naïve Bayes classification method is used to infer the user’s rating for aspects. it can be assumed that the overall rating on a product is weighted sum of the user’s specific rating on multiple aspects of the product, where the weights basically measure the degree of importance of the aspects. The thesis proposes an approach to infer the rating and weight of aspects. More specifically, the weight of an aspect is calculated by leveraging the aspect word frequency within the review and the aspect consistency across all reviews. The Fig. 2.3 summaries the three tasks mentioned above.

2.2 Related work

2.2.1 Aspect extraction

2.2.2 Aspect sentiment classification

2.2.3 Estimating aspect weight

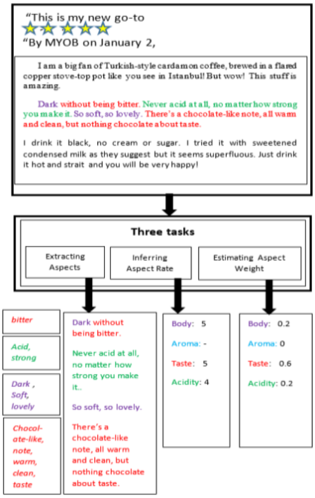

Figure 2.3 *Sub-tasks of the aspect-level sentiment analysis problem*

2.3 Problem definition

The review's text is denoted by d_i. Each review's text di can contain multiple sentences, each sentence contains many words w_j.

Definition 2.1 Review Text Documents: $\mathcal{D} = \{d_1, d_2, ..., d_D\}$ is a set of reviews about a product.

Definition 2.2 Vocabulary: Assume that, there are V words extracted from the set of reviews $\mathcal{D}$. Which is called a word dictionary $\mathcal{V} = \{w_j | j = \overline{1, V}\}$.

Definition 2.3 Aspect: An aspect is a feature (an attribute or a component) of a product. We assume that there are K aspects mentioned in all reviews, denoted by $\mathcal{A} = \{a_k | k = \overline{1, K}\}$. An aspect is represented by a set of words and denoted by $a_k = \{w | w \in V, A(w) = k\}$, where a_k is the name of the aspect, w is a word from the set $\mathcal{V}$ and $A(.)$ is an operator that maps a word to the aspect.

Definition 2.4 Aspect Core Words: Given an aspect a_k, a very small set of words in $\mathcal{V}$ describes the aspect very clearly a_k called aspect core word, denoted by $\mathcal{C}_k = \{w_{kj} \in \mathcal{V} | w_{kj} \rightarrow a_k, j = \overline{1, N}\}$, where w_{kj} is a word that describes the aspect a_k, N is the number of aspect core word a_k. The set of core word aspects does not belong to another aspect core set.

Definition 2.5 Aspect Words: Set of words in $\mathcal{V}$ which can describe aspect a_k (these words are different from aspect core words) called aspect words, denoted by $\mathcal{T}_k = \{w_{kj} \in \mathcal{V}, w_{kj} \notin \mathcal{C}_k | w_{kj} \rightarrow a_k, j = \overline{1, M}\}$. M is the number aspect word a_k.

Definition 2.6 Aspect Rating: Given a review d_i, a K-dimensional vector $\mathbf{r}_i \in \mathbb{R}^K$ is used to represent the scores of K aspects in the review text d_i, denoted by $\mathbf{r}_i = \{r_{i1}, r_{i2}, ..., r_{iK}\}$, where r_{ik} is a number indicating the user's opinion assessment on aspect a_k, and $r_{ik} \in [r_{min}, r_{max}]$ (e.g., the range of r_{ik} can be from 1 to 5).

Definition 2.7 Aspect Weight: Given a review d_i, a K-dimensional vector $\alpha_i \in \mathbb{R}^K$ is used to represent the user's interest in K aspects in the review text d_i, denoted by $\alpha_i = \{\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK}\}$, where α_{ik} is a number measuring the degree of importance of aspect a_k, and $\alpha_{ik} \in [0, 1]$, and $\sum_{k=1}^{K} \alpha_{ik} = 1$.

Definition 2.8 Review Overall Rating: Given a review d_i, a numeric $y_i \in \mathbb{R}^+$ represents the review's overall rating of a product across all aspects of the product. This overall score value is similar to the aspect rating score.

Extracting aspect: It is assumed that each aspect is a probability distribution over words and each sentence in a review's text can mention more than one aspect. The goal of this task is to extract aspects mentioned in a review.

Inferring aspect rate: This task is to infer the vector $\mathbf{r}_i$ of aspect ratings (definition 2.6) given a review d_i. Rating of an aspect reflects the user's sentiment on the aspect which is often expressed in positive or negative words.

Estimating aspect weight: This task is to estimate non-negative weights α_i that a user places on aspect a_{ik} of review d_i (definition 2.7). Essentially, the weight of an aspect measures the level of importance given by the user to that aspect.

2.4 Aspect-based sentiment analysis system for online product reviews

2.4.1 Aspect extraction using conditional probabilistic with Bootstrapping

The aspect label is determined based on the set of relevant words called aspect words or terms $\mathcal{T}_k$. Each aspect in the universal label set is provided with some initial core terms $\mathcal{C}_k$. It is assumed that the universal set of all possible aspects for each product is readily available. The aspect extraction task then becomes how to correctly assign existing aspects to sentences in the review. The main challenge here is that in many reviews, sentences do not contain enough core terms or even do not have any core term at all, and thus may be assigned with the wrong aspects. This problem is solved by repeatedly updating and enlarging the set of core terms to the set of aspect words by using the conditional probability technique combined with the bootstrap technique.

Suppose that $\mathcal{A} = \{a_1, a_2, ..., a_K\}$ is the set of K aspects. a_k is the set of words representing the aspect a_k and their occurrence frequency is always greater than the threshold θ. The goal is to determine the set of words that appear in the sentences of the entire corpus of the aspect a_k. The set of words of two aspects can overlap, such that some terms may belong to multiple aspects. First, sentences that contain at least one word in the original core terms C_k of the aspect are located. Then, all words including nouns, noun phrases, adjectives, and adverbs that appeared in these sentences are found. Words that occur more than a given threshold θ are inserted into the set of aspect words. Words with the maximum number of occurrences in the set of new-found aspect words are added to the set of core terms. The new set of aspect words with core terms excluded is used to find new sentences. The above-mentioned process is repeated until no more new words are found.

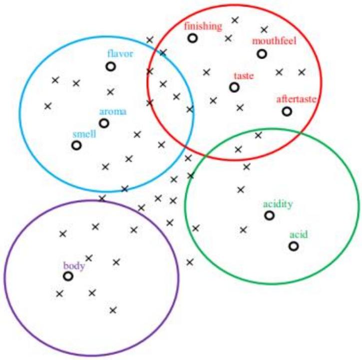

Figure 2.4 *Core terms with aspects*

2.4.2 Aspect score prediction based on Naïve Bayes multi-classification

The aspect rating problem is treated as the problem of multi-label classification, in which ratings (from 1 to 5) as considered as labels, and sentiment words are used as features. In addition, a number of bi-gram features are extracted according to fixed syntactic patterns.

Given a review's text d_i, the rating of an aspect a_k with Q features (denoted by f_q) are inferred based on the probability r_{ik} thuộc về lớp $c \in C_{class} = \{1, 2, 3, 4, 5\}$. The probability is as:

$$p(r_{i_k} \in c | f_1, f_2, ..., f_q) = \frac{p(f_1, f_2, ..., f_q | r_{i_k} \in c) * p(r_{i_k} \in c)}{p(f_1, f_2, ..., f_q)} \qquad (2.1)$$

It is assumed that the features are independent, the rating r_{ik} is the label c that maximizes $p(r_{ik} \in c | f_1, f_2, ..., f_Q)$.

$$\widehat{c} = \arg\max_{c \in C_{class}} \left(p(r_{i_k} \in c) * \prod_{q=1}^{Q} p(f_q | r_{i_k} \in c) \right) \qquad (2.4)$$

2.4.3 Estimate aspect weights based on aspect word frequency in the review and the corpus

For users, if an aspect is important, they will mention more about it in the review. Moreover, an idea that an important aspect is often shared by many users. The weight measure of aspect a_k in the review d_i is denoted by ED_{ik}, and the weight measurement of the aspect across all reviews is denoted by EC_k.

$$ED_{ik} = \frac{\sum_{j=1}^{N_i} w_{ikj}}{N_i} \tag{2.5}$$

in which: w_{ikj} is the j-th word in the aspect words of aspect a_k, and N_i is the number of aspect words that occur in the review's text d_i for all aspects.

$$EC_k = \frac{\sum_{h=1}^{M} s_{kh}}{M} \tag{2.6}$$

in which: s_{kh}, is the *k-th* sentence in the corpus labeled by the aspect a_k, and M is the number of all sentences in the corpus.

Finally, the weight α_{ik} for an aspect a_k of review d_i is calculated as:

$$\alpha_{ik} = \frac{ED_{ik} * EC_k}{\sum_{k=1}^{K} ED_{ik} * EC_k} \tag{2.7}$$

2.5 Experimental results

2.5.1 Data set

The experiments were performed on three datasets hotel reviews collected from Tripadvisor.com, beer reviews collected from Beeradvocate.com, and Trung Nguyen coffee reviews collected from the Amazon.com website.

2.5.2 Preprocessing and feature selection

2.5.3 Results and discussion

To evaluate the performance, the precision measure is used. Table 2.5 shows the performance of the proposed method in the aspect extraction problem. Our method yields up to average precision of 0.786, 0.803, and 0.653 for the hotel data set, beer data set and coffee data set, respectively. The proposed method achieves good performance on hotel and beer datasets. However, for the coffee dataset, the results are not as good as expected.

We then search for the best threshold θ at which our method performs the best. The results are shown in Fig. 2.5, where the threshold θ of about 0.15 is the best one.

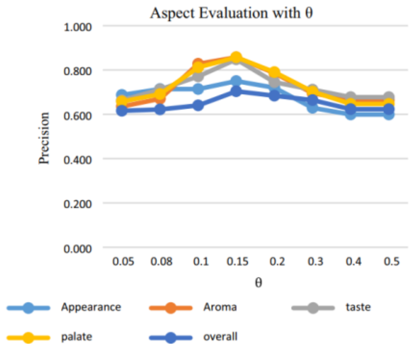

Figure 2.5 *Aspect evaluation with θ*

Table 2.5 *Aspect extraction results on three datasets Hotel, Beer, Coffee*

Hotel Data set		Beer Data set		Coffee Data set	
Aspect	$P_{recision}$	Aspect	$P_{recision}$	Aspect	$P_{recision}$
Value	0.747	Appearance	0.750	Aroma	0.667
Room	0.837	Aroma	0.857	Taste	0.677
Location	0.814	Palate	0.857	Acidity	0.667
Cleanliness	0.764	Taste	0.848	Body	0.600
Check in/front desk	0.850	Overall	0.704		
Service	0.754				
Business service	0.737				
Average	0.786	Average	0.803	Average	0.653

The proposed method of the thesis is compared with the frequency-based method with the hotel data set. Figure 2.9 that our method outperforms Long's method in the room (R), service (S), and cleanliness (C) aspects. But Long's method outperforms us in detecting the value (V) aspect.

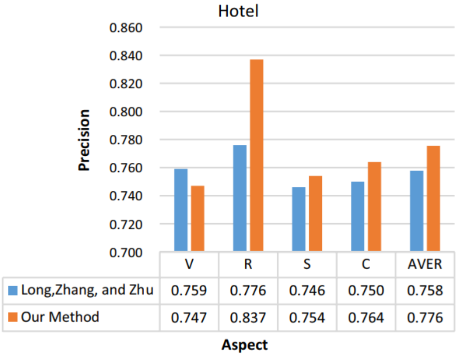

Figure 2.9 *The results of our method and Long et al. method*

The proposed method is also compared with two methods based on the topic model in (Semi-supervised and supervised PALELAGER) and in (LDA) with the beer dataset. Figure 2.5 shows that our method outperforms LDA by a large margin, and slightly outperforms PALE LAGER (a semi-supervised method) and PALE LAGER (a supervised method).

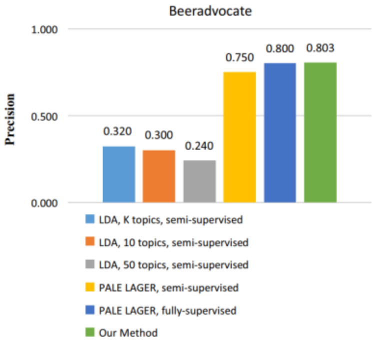

Figure 2.5 *The results of our method and LDA, PALE LAGER*

Aspect ranking prediction

The mean square error measure (named $\Delta^2{}_{aspect}$), the aspect correlation measure (named ρ_{aspect}), and *aspect correlation across reviews* measure (named ρ_{review}) are used. The results of the proposed method are compared with the two methods of Long et al., and Wang's method on the hotel data set. The comparison results are shown in Table 2.9.

Table 2.9 *Comparison with other models for referring aspect ratings*

Method	Δ^2_{aspect}	ρ_{aspect}	ρ_{review}
Long et al. with SVM	0.286	0.557	0.708
Long et al. with BN	0.441	0.429	0.591
LRR	0.896	0.464	0.618
Our method	0.101	0.583	0.757

Estimating aspect weight

Our method is compared with Wang's method based on the mean square errors of overall rating ($\Delta^2{}_{overallrating}$) for the three data sets. The results are shown in Table 2.10.

Table 2.10 *MSE of overall rating prediction*

Method	Product datasets		
	Hotel	Beer	Coffee
LRR	0.905	0.856	1.234
The proposed method	0.1456	0.1423	0.1904

2.6 Conclusion of Chapter 2

In Chapter 2, the Ph.D. student presents a model that solves three important tasks of the aspect-based sentiment analysis problem: (1)

extracting aspects mentioned in the review of a product using the conditional probability of words combined with the Bootstrapping algorithm; (2) infer a user rating for each defined aspect based on the Naive Bayes classifier; (3) estimate the weight that users place on each aspect using the number of occurrences of words discussing that aspect in a review and the frequency of sentences discussing the same aspect edge on all reviews.

CHAPTER 3 ASPECT EXTRACTION BASED ON WORD2VEC COMBINED LANGUAGE MODEL

3.1 Introduction to the problem

3.2 Related word

3.3 Some basic concepts in the aspect extraction model based on Word2vec

Definition 3.1 Word vector: Given a word w_j a vector P dimensional $\mathbf{x}_{wj} \in \mathbb{R}^P$ is used to represent P different contexts of word w_j in the entire context space of the corpus. The notation $\mathbf{x}_{wj} = \{x_{1wj}, x_{2wj}, \ldots, x_{pwj}\}$, where x_{pwj} is a real numeric value obtained by training Word2vec.

Definition 3.2 Aspect core word vector: Each core word of aspect a_k, $w_{kj} \in \mathcal{C}_k$ is mapped respectively to a vector in the set of word vectors called core word Vector and has the notation $\mathbf{x}_{coreak}$.

Definition 3.3 *supp*($w_j \rightarrow a_k$): The support of the word w_j for the aspect a_k is a value representing the probability that the word w_j can describe the aspect a_k. The support is calculated based on the improvement of the Euclidean measure as shown in Equation (3.1).

$$supp(w_j \longrightarrow a_k) = \frac{1}{N} \sum_{t=1}^{N} \frac{1}{\sum_{p=1}^{P} (x_{pw_j} - x_{pcore_{ta_k}})^2} \qquad (3.1)$$

in which: supp($w_j \rightarrow a_k$) supporting of w_j to aspect a_k, N is the number of core terms for aspect a_k, P is the number of dimensions of a word, x_{pwj} value of dimension p of word w_j; $x_{pcoretak}$ dimension p of core term *t-th* that belongs to aspect a_k.

Definition 3.4 *supp*($S \rightarrow a_k$): The support of a sentence S for aspect a_k is a value representative of the probability that sentence S can describe aspect a_k. The support of sentence S for aspect a_k is calculated based on the average support of all words w_j in sentence S for aspect a_k according to the formula (3.2).

$$supp(S \longrightarrow a_k) = \frac{1}{Q} \sum_{j=1}^{Q} supp(w_j \longrightarrow a_k) \qquad (3.2)$$

in which: *supp*($S \rightarrow a_k$) supporting of sentence S to aspect a_k; *supp*($w_j \rightarrow a_k$) supporting of word w_j to aspect a_k; Q number of words in sentence S.

3.4 Aspect extraction model based on word2vec and support score

Each aspect a_k is represented by a set of words. Words are represented as Word2vec so that we can consider words in different contexts to improve the accuracy of the problem. The proposed model is shown in Figure 3.2.

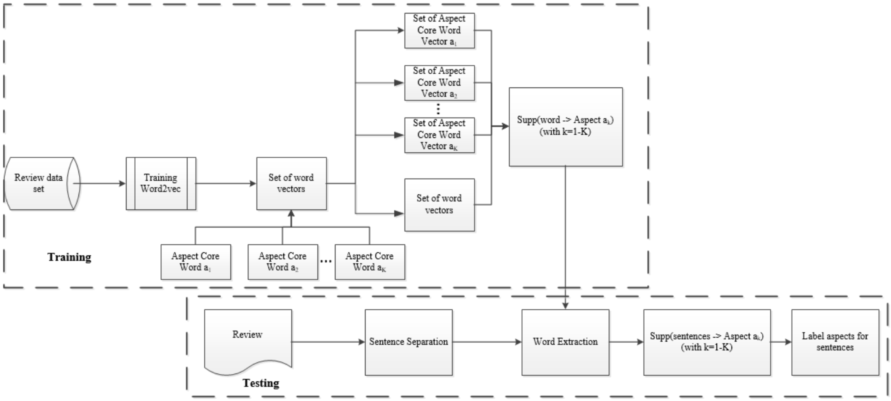

Figure 3.4 *Aspect labeling of sentences based on Word2vec and support score*

Training phase:

Step 1 (Data): separates sentences, and normalizes sentences.

Step 2 (Training with Word2Vec): use the word2vec tool running on python language to vectorize words.

Step 3 (Building core terms for each aspect): We perform to classify and assign aspects to words (we called aspect core terms).

Step 4 (Calculating supp(word → aspect)): The supporting measure of word w_j is calculated in aspect a_k like Equation (3.1)

Testing phase:

Step 1 (Sentence segmentation): separates sentences, and normalizes sentences.

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

Step 3 (supp(sentence -> aspect)): Based on the supporting measure of words to aspect, we can calculate the supporting of a sentence to aspect by the formula (3.2)

Step 4 (assign aspect labels to sentences): compare the support of the sentence with the threshold or take the maximum value to determine the aspect label for the sentence.

3.5 Experimental results

To evaluate the effectiveness of the proposed method, in this part, the thesis uses two measures, precision and recall. The test results on three data sets are shown in tables 3.3, 3.4, 3.5.

The proposed method is tested and compared with two basic methods, LDA and Long et al., with the hotel data set using the precision measure. The results are shown in Table 3.6.

Table 3.3 *Aspect extraction results for the Hotel dataset*

Aspect	Precision	Recall
Value	0.774	0.753
Room	0.788	0.751
Location	0.823	0.794
Cleanliness	0.767	0.728
Check in/ front desk	0.804	0.800
Service	0.736	0.684
Business service	0.850	0.835
Average	**0.792**	**0.764**

Table 3.4 *Aspect extraction results for the Beer dataset*

Aspect	Precision	Recall
Appearance	0.795	0.800
Aroma	0.875	0.901
Palate	0.862	0.792
Taste	0.843	0.826
Overall	0.821	0.803
Average	**0.839**	**0.824**

Table 3.5 *Aspect extraction results for the Coffee dataset*

Aspect	Precision	Recall
Aroma	0.702	0.684
Taste	0.666	0.659
Acidity	0.654	0.600
Body	0.712	0.720
Average	**0.684**	**0.666**

Table 3.4 *Compare the results of the proposed method with the LDA method and Long et al. on the Hotel dataset with measure precision*

Aspect	PP LDA [26]	PP Long et al. [100]	Proposed method
Value	0.65	0.76	0.77
Room	0.47	0.78	0.79
Location	0.56	0.75	0.82
Cleanliness	0.60	0.75	0.77
Check in/front desk	0.65	0.74	0.80
Service	0.59	0.75	0.74
Business service	0.60	0.75	0.85
Average	**0.589**	**0.754**	**0.791**

3.6 Conclusion of Chapter 3

In this chapter, the Ph.D. student proposed an aspect extraction model based on efficiently exploiting feature representations from vector form and using them to calculate the weights of core terms using a support measure. This method works well on real-world datasets and it can be applied to a number of different fields.

CHAPTER 4: ASPECT SENTIMENT CLASSIFICATION BY COMBINATION BASE CLASSIFIERS

4.1 Introduction to the problem

We predict customer reviews belonging to one of the following five classes based on the emotional and rational evaluation. The main challenge in inference rating is how to correctly classify a review into nearby classes due to relatively small differences in evaluation scores, high uncertainty, high ambiguity misclassification errors that often occur, or when the feature vector does not contain sufficient information to identify an exact class or some classes may have similar probability scores. Another important difficulty is the problem of imbalanced data. The third challenge is that sparse data which highly reliant on context and usually does not provide enough background information for a good discrimination function between different texts.

To overcome the above difficulties, Our idea is to combine different classifiers, that can be complementary to one another, overcoming each classifier's weaknesses, and providing multifarious types of pieces of evidence, improvement in the classification accuracy can be obtained, especially in cases of high uncertainty and ambiguity.

The Ph.D. student proposes to use a method based on Dempster-Shafer (DS) theory and use only two powerful classifiers, multi-class SVM and a multi-class classification algorithm based on the prototypical disjunction interaction model (or OR Gate Bayesian Network (OGBN)). The goal of the proposal: use the fewest classifiers, solve the problem of imbalanced data, and improve the performance of multiclass classification.

Text is preprocessed, the features are uni-gram, bi-gram, information gain (IG), and mutual information (MI). The reviews are classified based on SVM and OGBN. The output of the SVM algorithm is passed through a conversion function to a probability value. The corresponding probability value of the SVM together with the output probability OGBN becomes the input of the combination rule DS. The final rating of the review is the class whose true value is the greatest.

4.2 Related work

4.3 Multiclass classification using Bayesian Noisy OR-gate for sentiment classification

As introduced in Section 4.1, the thesis proposes a combination model that combines output probabilities from two basic classification algorithms (SVM and OGBN) based on DS association rules. The final prediction class is based on the aggregated results from the underlying algorithms. Figure 4.1 depicts the process in the proposed model.

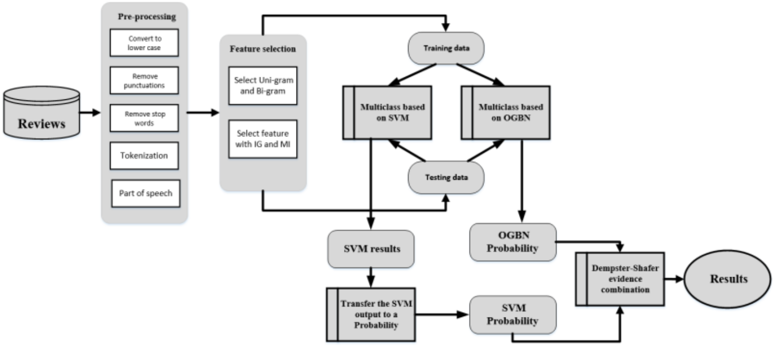

Figure 4.1 *Sentiment multi-class classification by combining SVM and OGBN based on Dempster-Shafer's rule*

The text is preprocessed (see Section 2.5.2), and the uni-gram, bi-gram, IG, and MI features are selected. That features become the input to the basic classification algorithms SVM and OGBN. The output of the SVM algorithm is a function of the distance from the data point to the hyperplanes, which is not a probability value. Therefore, this distance score is passed through a conversion function to a probability value. The corresponding probability value of the SVM together with the output probability OGBN becomes the input of the association rule DS. The final rating of the assessment is the class whose combined validation value is the largest.

4.3.1 Multi-class classification based on SVM

With a multiclass classification problem, a One-vs-all (OVA) strategy is chosen. There are a total of n binary classifiers needed to train. A new sample *x* is assigned to the class that its' classifier outputs (4.1) the maximum positive value (i.e., maximal margin) as in (4.2).

$$y(x) = \mathbf{wx} + b = \sum_{i=1}^{D} \alpha_i c_i (\mathbf{x} x_i + b) \tag{4.1}$$

$$c = arg \max_{1 \leq c \leq C} y_c(x) \tag{4.2}$$

4.3.2 Convert SVM outputs to posterior probabilities

As we will use the Dempster-Shafer method to combine classifiers, we need to calibrate the multi-class SVM classifier to output posterior probability values. Platt provided a kernel method that fits a sigmoid function that maps SVM outputs to posterior probabilities as follows:

$$p(c = 1|y(x^*)) = \frac{1}{1 + exp(Ay(x^*) + B)} \tag{4.3}$$

in which *f*(*x*) is defined in (4.11).

For estimating the parameters A and B of (4.3) fitting the sigmoid, the pseudocode in Platt's proposal is used.

4.3.3 Sentiment multiclass Classification based on OR Gate Bayesian Network

The OR Gate Bayesian Network method inherits the advantages of Bayesian networks. The method is naturally suitable for the multi-class problem and works well on highly unbalanced data, it also reduces the computational complexity compared to Bayesian networks from $O(2^n)$ to $O(n)$.

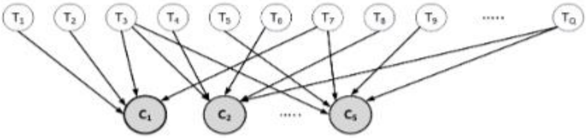

Figure 4.5 *OR gate Bayesian network classification*

The OGBN model for sentiment rating is constructed as follows: a set of features $\{f_q\}$, each feature $\{f_q\}$ is associated with a binary variable T_q also called a *cause node*; There are C_{class} *effect nodes corresponding to* C_{class} classes. The network structure is fixed. Having an arc going from each cause node T_q to the effect node C_j if feature $\{f_q\}$ appears in training data of class c_j. The posterior probability of each class, c_j given review d_i is computed as follows:

$$p(c_j|d_i) = 1 - \prod_{T_q \in P_a(c_j) \cap d_i} (1 - p(f_q)) \tag{4.7}$$

This probability can be estimated directly $\hat{p}(c_j|f_q)$ using Laplace estimation:

$$\widehat{p}(c_j|f_q) = \frac{N_{jq} + 1}{N_{\bullet q} + 2} \tag{4.8}$$

where N_{jq} is the number of times that f_q appears in documents of class c_j;

$N_{\bullet q}$ is the number of times f_q appears in all documents, i.e. $N_{\bullet q} = \sum_{c_j} N_{jq}$. Then the classification function of d_i will be:

$$c = arg \max_{c \in \{c_1, c_2, \dots, c_5\}} (1 - \prod_{T_q \in Pa(C_j) \cap d_i} (1 - p(f_q)) \tag{4.9}$$

4.3.4 Dempster's Rule for Classifier Combination

The power set $\mathbf{P}(C)$ is the set of all possible subsets of classes, $\mathbf{P}(C) = \{\emptyset, \{c_1\}, \dots, \{c_n\}, \{c_1, c_2\}, \dots, \{c_1, \dots, c_n\}\}$. The example, $n = 5$ then $\mathbf{P}(C)$ will be $2^5 = 32$ subsets. The DS theory assigns a *mass value m* between 0 and 1 to each subset $A \in \mathbf{P}(C)$ of the power set and satisfies the following:

$$m(\phi)=0;\ \sum_{A\in\mathbf{P}(C)} m(A)=1 \tag{4.10}$$

Given two pieces of evidence represented by two basic probability assignments, m_1 and m_2, Dempster's rule of combination (also called the orthogonal sum mass function and denoted by $m = m_1 \oplus m_2$) is defined as follows:

$$m_{1,2}(A)=m_1\oplus m_2(A)=\frac{\sum_{X,Y\in\mathbf{P}(C);X\cap Y=A} m_1(X)m_2(Y)}{1-\sum_{X,Y\in\mathbf{P}(C);X\cap Y=\emptyset} m_1(X)m_2(Y)} \tag{4.11}$$

for $A \in \mathbf{P}(C)$ called a hypothesis.

Let $\Theta = \mathbf{P}(C)/C$, that means Θ accounts for all those subsets of $\mathbf{P}(C)$ having cardinality greater than one, for each hypothesis corresponding to a singleton class c_j, we have:

$$\sum_{X,Y\in\mathbf{P}(C);X\cap Y=c_j} m_1(X)m_2(Y)\approx m_1(c_j)m_2(c_j)+m_1(c_j)m_2(\Theta)+m_1(\Theta)m_2(c_j) \tag{4.12}$$

According to (4.10), $m(\Theta)$ is given by:

$$m(\Theta)=1-\sum_{c_j\in C} m(c_j) \tag{4.13}$$

Note that $m(\Theta)$ in (4.13) still accounts for subsets of $\mathbf{P}(C)$ that are not superset of c_j, To reduce the impact of such subsets, we use the following $\tilde{m}$:

$$\tilde{m}(\Theta)=\frac{|\{A\in\Theta;A\ni c_j\}|}{|\Theta|}\left(1-\sum_{c_j\in C} m(c_j)\right) \tag{4.14}$$

To construct the mass function for each class c_j from a sample text review d_i, we rely on the confusion matrix (CM_φ) and the probability value of each class c_j ($p(c_j|d_i)$) determined by a classifier φ, mass function for the class c_j provided by the classifier φ as follows:

$$m_\varphi(c_j)=\frac{2P_\varphi(c_j)R_\varphi(c_j)}{P_\varphi(c_j)+R_\varphi(c_j)}\cdot\frac{p_\varphi(c_j|d_i)}{\sum_{j=1}^{n} p_\varphi(c_j|d_i)} \tag{4.15}$$

4.4 Experiment and Results

4.4.1 Data sets

The review distribution of five classes in the three datasets is shown in Table 4.5.

__Table 4.5__ The distribution of emotional classes in the Data set

Dataset		class emotional negative	class rational negative	class neutral	class rational positive	class emotional positive	sum
Reviews	Hotel	12,565	13,415	24,892	61,254	81,535	193,661
	Beer	230	1,245	5,785	27,224	15,516	50,000
	Coffee	654	857	1,413	4,142	4934	12,000

4.4.2 Preprocessing and feature engineering

__Table 4.6__ Dimensions of two feature sets in three datasets

The dataset	The number of Uni+Bi feature	The number of Uni+Bi+IG+MI feature
Hotel	69,314	6,000
Beer	55,231	5,000
Coffee	19.099	2,000

The thesis builds two sets of features: the basic feature set (uni-gram, bi-gram); The reduced feature set based on the base feature set through feature filters (IG combined with MI). Three experiments were performed to evaluate the performance of the proposed method.

4.4.3 Kết quả và thảo luận

The first experiment compares the performance of an SVM-based, Bayesian Noisy OR-gate multiclass classifier using two different input feature sets. Table 4.7 shows the performance of the SVM and OGBN-based method on three datasets. The OGBN-based classifier performed better than the SVM-based classifier with OVA in all datasets. This result confirms our previous analysis that SVM performs well with binary text classification, but has difficulty dealing with multi-text classifiers. The OGBN-based classifier performs better for feature sets with large dimensionality ("Uni+Bi"), while the SVM-based method performs well with reduced feature sets ("Uni+Bi+ IG+MI”).

The second test is to evaluate the effect of combining two basic classifiers using DS theory. We will evaluate the overall improvement of the association model, and evaluate the problem of imbalanced data and the problem of misclassification between neighboring classes.

Table 4.8 shows that the DS-based association method performed better than both SVM-based and OGBN-based methods for all three datasets. The results show that the combined method is slightly superior to the SVM-based classifier (ACC from 3.27% to 5.75%) and to the OGBN-based classifier (ACC from 1.82% to 2.54%) the results were covered by the majority of layers.

Table 4.7 Comparison of two base classifiers on three datasets

Dataset	The metric					
	classifier	feature	P(%)	R(%)	F1(%)	Accuracy(%)
Beer	SVM-based	Uni+Bi	74.37	79.42	76.81	89.54
		Uni+Bi+IG+MI	78.13	83.44	80.70	91.36
	MCDI	Uni+Bi	82.29	92.18	86.95	93.96
		Uni+Bi+IG+MI	83.11	91.35	87.03	93.54
Hotel	SVM-based	Uni+Bi	86.43	86.45	86.44	86.43
		Uni+Bi+IG+MI	87.75	89.36	88.55	90.39
	MCDI	Uni+Bi	89.06	90.80	89.92	91.45
		Uni+Bi+IG+MI	88.62	90.21	89.41	91.12
Coffee	SVM-based	Uni+Bi	81.40	81.82	81.61	82.83
		Uni+Bi+IG+MI	89.33	89.41	89.37	90.08
	MCDI	Uni+Bi	94.41	93.42	93.91	94.08
		Uni+Bi+IG+MI	93.77	92.95	93.36	93.67

Table 4.8 Comparison of the combined method with two base classifiers

Dataset	The metric				
	classifier	P(%)	R(%)	F1(%)	Accuracy(%)
Beer	SVM-based	78.13	83.44	80.70	91.36
	MCDI	83.11	91.35	87.03	93.54
	DS	**88.17**	**94.69**	**91.32**	**95.36**
Hotel	SVM-based	87.75	89.36	88.55	90.39
	MCDI	88.62	90.21	89.41	91.12
	DS	**91.89**	**92.76**	**92.32**	**93.66**
Coffee	SVM-based	89.33	89.41	89.37	90.08
	MCDI	93.77	92.95	93.36	93.67
	DS	**95.81**	**95.63**	**95.72**	**95.83**

Tables 4.9, 4.10, 4.11 present the number of samples that are misclassified between two adjacent classes according to three methods, the classes of emotional negative, rational negative, neutral, rational positive, and emotional positive are denoted by c_1, c_2, c_3, c_4, c_5 respectively.

Table 4.9 Mis-classified samples of adjacent classes of three models on the Beer Dataset.

	Classifier	SVM based	MCDI	DS combined
The number of mis-classified samples	$c_1 \to c_2$	6	2	2
	$c_2 \to c_1$	10	0	3
	Total	16	2	5
	$c_2 \to c_3$	10	7	2
	$c_3 \to c_2$	36	34	19
	Total	46	41	21
	$c_3 \to c_4$	18	14	7
	$c_4 \to c_3$	56	29	29
	Total	74	43	36
	$c_4 \to c_5$	132	78	78
	$c_5 \to c_4$	51	45	37
	Total	183	123	115

Table 4.10 Mis-classified samples of adjacent classes of three models on the Hotel Dataset.

	Classifier	SVM based	MCDI	DS combined
The number of mis-classified samples	$c_1 \to c_2$	114	100	56
	$c_2 \to c_1$	63	56	53
	Total	177	156	109
	$c_2 \to c_3$	27	25	25
	$c_3 \to c_2$	101	95	68
	Total	128	120	93
	$c_3 \to c_4$	104	104	100
	$c_4 \to c_3$	136	129	122
	Total	240	233	222
	$c_4 \to c_5$	241	232	180
	$c_5 \to c_4$	312	262	163
	Total	553	494	343

Table 4.11 Mis-classified samples of adjacent classes of three models on the Coffee Dataset.

	Classifier	SVM based	MCDI	DS combined
The number of mis-classified samples	$c_1 \to c_2$	18	16	6
	$c_2 \to c_1$	10	4	4
	Total	28	20	10
	$c_2 \to c_3$	8	8	8
	$c_3 \to c_2$	7	4	4
	Total	15	12	12
	$c_3 \to c_4$	4	4	4
	$c_4 \to c_3$	7	4	4
	Total	11	8	8
	$c_4 \to c_5$	23	12	8
	$c_5 \to c_4$	18	16	7
	Total	41	28	15

4.5 Conclusion of chapter 4

In this chapter, the thesis considers solving the task of classifying sentiment of multi-class aspects. The Ph.D. student proposed a powerful associative model to solve the above problem using a method based on Dempster-Shafer's theory with a careful selection of the best complementary base classifiers for each other. By applying the analysis of the strengths and weaknesses of the existing methods, the Ph.D. student has come up with two candidates for the combined classifier, namely SVM-based and multiclassification methods. type based on Bayesian OR gate network. The results show the superior effectiveness of the combined method compared with the two basic methods. The results also show the ability to overcome the problems of unbalanced data, the ambiguity of data and the adjacency of neighboring classes.

CONCLUSION AND DEVELOPMENT DIRECTION

1. The results of the thesis

Aspect-based sentiment analysis of online product reviews is considered a useful tool to explore user personalization, predict consumer trends, and product market orientation. Research in this thesis develops a number of machine learning algorithms to improve the quality of mining and analysis of facets. Some conclusions are as follows:

- The thesis proposes a system to perform three tasks: aspect extraction, aspect sentiment score prediction, and aspect weight estimation of the aspect-based opinion analysis problem. ***Aspect extraction task***, The thesis proposes a semi-supervised learning technique based on conditional probability combined with bootstrapping algorithm to perform the problem. The proposed method can solve labeled data problems, implicit aspect detection, and low-frequency aspects. ***Aspect sentiment score prediction***, the Naive Bayes supervised learning method is implemented. This approach is capable of solving multiclass problems and imbalanced data. ***Aspect weights estimation***, the unsupervised approach based on the content of user reviews and the universality of the entire studied corpus. The proposed method helps to solve the personalization of each user but does not require the rating of each aspect or the overall rating of the article.
- The thesis proposes a semi-supervised method to improve aspect extraction performance based on W2V representation combined with a language model. The proposed method can solve well for hidden aspect extraction and especially solve the problem of context dependence of words in this task.
- The thesis proposes a method that combines two powerful classifiers, Support Vector Machine and OR Gate Bayesian Network, based on Dempster's theory to solve the facet emotion classification task. The proposed method has superior performance compared to the two basic methods. In particular, the combined method can solve the problem of separating close classes, and the problem of unbalanced data in the multi-class classification problem.

2. Development Orientation

From the research results that have been carried out and the limitations pointed out, the researcher proposes a number of expansion studies as follows:

- First, carry out research to synthesize views from the published results of the thesis.

- Second, expand the scope of research on opinion reviews other than product review articles on online media.
- Third, further study machine learning methods so that different learning methods can be combined to improve the overall performance of the system in the given task.

LIST OF WORKS OF THE AUTHOR

1. Nguyen Thi Ngoc Tu, Nguyen Thi Thu Ha, Nguyen Long Giang, Nguyen Viet Anh, Nguyen Tran Quoc Vinh. "An effective algorithm multi-class classification of online product reviews". The national conference on Fundamental and Applied IT research, Ha Noi, 11/2022, ISBN: 978-604–357-119-6 p517-526.
2. Tu Nguyen Thi Ngoc, Ha Nguyen Thi Thu, Viet Anh Nguyen. "*Language model combined with word2vec for product's aspect based extraction*". ICIC Express Letters, Volume 14, Number 11, 2020 ISSN 1881-803X p1033-1040 (SCOPUS).
3. Tu Nguyen Thi Ngoc, Ha Nguyen Thi Thu, Viet Anh Nguyen. "*Mining Aspects of Customer's Review on the Social Network* ". Journal of Big Data, Volume6, Issue 1, 12/2019 ISSN: 2196-1115 (SCOPUS - Q1).
4. Nguyen Thi Ngoc Tu, Bui Khanh Linh, Nguyen Thi Thu Ha, Nguyen Viet Anh, Nguyen Ngoc Cuong. "*Extract aspect of product based on language model combined with Word2Vec*". National Symposium of Selected ICT Problems, Thanh Hoa, 27-28/7/2018, ISBN: 978-604-67- 1104-9 p343 - 349.
5. Nguyen Thi Ngoc Tu, Nguyen Duc Long, Nguyen Khac Giao, Nguyen Thi Thu Ha, Nguyen Viet Anh. "*A method of opinion mining by consumers for product quality based on reviews*". The national conference on Fundamental and Applied IT research, Da Nang, 8/2017, ISBN: 978-604–913-614-6 p585-594.