Life-long learning for cross-domain Vietnamese sentiment classification

Quang-Vinh Ha, Bao-Dai Nguyen-Hoang, and Minh-Quoc Nghiem

Faculty of Information Technology
University of Science, VNUHCM
227 Nguyen Van Cu, Dist. 5, Ho Chi Minh city, Vietnam
{1212304,1212069}@student.hcmus.edu.vn,nqminh@fit.hcmus.edu.vn

Abstract. This paper proposes an improvement to life-long learning for cross-domain sentiment classification. Life-long learning is to retain knowledge from past learning tasks to improve the learning task on a new domain. In this paper, we will discuss how bigram and bag-of-bigram features integrated into a life-long learning system can help improve the performance of sentiment classification on both Vietnamese and English. Also, pre-processing techniques specifically for our cross-domain Vietnamese dataset will be discussed. Experimental results show that our method archives improvements over prior systems and its potential for cross-domain sentiment classification.

Keywords: sentiment classification; Vietnamese; supervised learning, life-long learning

1 Introduction

The rapid growth of e-commerce and the Web age quickly make the sentiment knowledge become an advantage to contribute more values to market predictions. Sentiment analysis remains a popular topic for research and developing sentiment-aware applications [1]. Sentiment classification, which is a subproblem of sentiment analysis task, is the task of classifying whether an evaluative text is expressing a positive, negative or neutral sentiment. In this paper, we focus on document-level binary sentiment classification, in which the sentiment is either positive or negative.

In recent years, most studies on sentiment classification adopt machine learning and statistical approaches [2]. Such approaches hardly perform well on real-life data, which contains opinionated documents from domains different from the domain used to train the classifier. To overcome this limitation, life-long learning [3], transfer learning [4], self-taught learning [5] and other domain adaptation techniques [4] were proposed. All mentioned methods is to transfer the knowledge gained from source domains to improve the learning task on the target domain.

Chen et al. [3] proposed a novel approach of life-long learning for sentiment classification, which is based on Naïve Bayesian framework and stochastic gradient descent. Although this approach could deal with cross-domain sentiment

classification, it used the "bag-of-words" model and faces difficulties when represent the relationship between words. For example, the phrase "have to", which is a common phrase in negative text (but much less important in positive text), cannot be taken advantage of with bag-of-words feature. This is especially true in isolated languages, such as Vietnamese, where words are not separated by white spaces.

As a resource-poor language, Vietnamese has quite few accomplishments in the field of sentiment classification. To the best of our knowledge, there is no study on Vietnamese cross-domain sentiment classification. There is also no suitable dataset with a reasonable amount of reviews and variance of products to apply life-long learning on Vietnamese.

In this paper, we propose the use of bigram feature to life-long learning approach on sentiment classification. Wang and Manning [6] proved that adding bigrams improves sentiment classification performance because they can capture modified verbs and nouns. We also created a dataset for Vietnamese cross-domain sentiment classification by collecting more than 15,000 reviews from the e-commerce website Tiki.vn [1] with 17 distinctive domains. We poposed combining the bigram feature with the Naïve Bayesian optimization framework. The proposed method has leveraged the phrases that contain sentiment better than that of Chen et al. [3] and outperforms other methods in both Vietnamese and English datasets.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the background and related work. Section 3 presents our method including how we add bigram and bag-of-bigram features to the life-long learning, and how we processed the raw reviews of the Vietnamese dataset to improve the performance. Section 4 describes the experimental setup and results. Section 5 concludes the paper and points to avenues for future work.

2 Related Work

Our work is related to life-long learning, multi-task learning, transfer learning and domain adaptation. Chen and Liu has exploited different types of knowledge for life-long learning on mining topics in documents and topic modeling [7, 8]. Chen and Liu [3], in their other work, also proposed the first life-long learning approach for sentiment classification. Likewise, Ruvolo and Eaton [9] developed a method for online multi-task learning in the lifelong learning setting, which maintains a sparsely shared basis for all task models. About domain adaptation, most of the work can be divided into two groups: supervised (Finkel and Manning 2009 [10], Chen et al. 2011 [11]) and semi-supervised (Kumar et al. 2010 [12], Huang and Yates 2010 [13]).

There are also many previous work on transfer learning and domain adaptation for sentiment classification. Yang et al. [14] proposed an approach based on feature-selection for cross-domain sentence-level classification. Other approaches

[1] http://tiki.vn/

include structural correspondence learning (Blitzer et al. [15]), spectral feature alignment algorithm (Pan et al. 2010 [16]), CLF (Li and Zong 2008 [17]). Similar methods can be found in the work of Liu [2].

In the field of sentiment analysis for Vietnamese, Duyen et al. [18] has published an empirical study which compared the use of Naïve Bayes, MEM and SVM with hotel reviews. Also using the corpus from Duyen, Bach et al. [19] proposed the use of user-ratings for the task. Term feature selection approach was investigated by Tran et al. 2011 [20], while Kieu and Pham [21] investigated a rule-based system for Vietnamese sentiment classification. As that being said, to the best of our knowledge, there is no previous work on domain adaptation or life-long learning as well as a appropriate dataset for Vietnamese (with a reasonable amount of reviews and variance of products).

3 Our Proposed Method

In this section, we describe our system for sentiment classification in a life-long learning setting, which is a combination of components to analyze reviews from many domains. The system takes customer reviews, from multiple types of products, as source domains. Each review can contain multiple sentences and it is labeled positive, negative or neutral based on how users rated them. From the source domains mentioned above, the system gains knowledge valuable to the learning task on target domain. Such knowledge is used to optimize the classifier on the target domain using stochastic gradient descent (SGD).

3.1 Overview of life-long learning for sentiment classification

As described in figure 1, the system contains three main modules: knowledge storing, optimization, and sentiment classification.

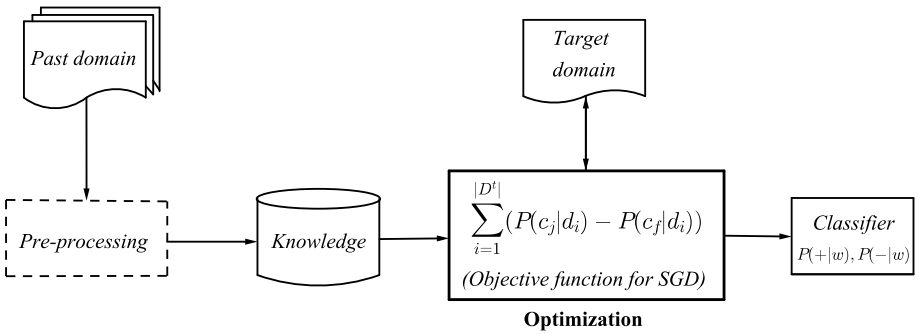

Fig. 1: Life-long learning for sentiment classification

Knowledge storing The system extracts the knowledge from the past domains, which is used to optimize the classifier on the target domain. There are three types of knowledge, including:

- The Prior probability $P_+^t(w|c)$ and $P_-^t(w|c)$ of each word, where t is a past learning task.
- Number of times a word appears in positive or negative in learning task: $N_{+,w}^t$, $N_{-,w}^t$. Similarly, the number of occurrences of w in the positive and negative documents are respectively $N_{+,w}^{KB} = \sum N_+^t$ and $N_{-,w}^{KB} = \sum N_-^t$
- Number of past tasks in which $P_{w|+} > P_{w|-}$ or vice versa: $M_{+,w}^{KB}$, $M_{-,w}^{KB}$. The two figures are used to leverage domain knowledge via a penalty term to penalize the words that appear in just a few domains.

Optimization With the help of all three types of knowledge mentioned above, this component is used to optimize the objective function on the training set of the target domain. The objective function is $\sum_{i=1}^{|D_i|} P(c_j|d_i) - P(c_f|d_i)$, in which c_j is the actual labeled class, c_f is the wrong class of the document d_i. We follow the SGD with similar regularization techniques proposed by Chen et al. [3]. Our optimized variables are $X_{+,w}$ and $X_{-,w}$, which are the occurrences of a word w in a positive and negative class, respectively. The objective function is optimized on each document of the target domain until convergence. After SGD, we use Bayes formula (see equation 1, 2) to create a classifier optimized for the target domain. Note that Laplace smoothing is applied in both cases.

$$P(+|w) = \frac{\lambda + X_{+,w}}{\lambda|V| + \sum_{v=1}^{V} X_{+,v}} \tag{1}$$

$$P(-|w) = \frac{\lambda + X_{-,w}}{\lambda|V| + \sum_{v=1}^{V} X_{-,v}} \tag{2}$$

Sentiment classification With the classifier optimized for the target domain, the system does sentiment classification task on each document of the test domain. Although the approach still follows Naïve Bayes framework, the way we classify differentiates between unigrams, bigrams, and bag-of-bigrams.

3.2 Bigrams

We propose the use of bigram feature, instead of unigram, on this type of sentiment classification. Wang and Manning [6] has proved that using bigram always improve the performance on sentiment classification. For instance, phrases such as "have to" in English or "không thích" (dislike) in Vietnamese can express sentiment well in the documents. These noun phrases and verb phrases cannot be captured by using unigram feature alone.

The way we integrate bigram feature into Naïve Bayesian framework for lifelong learning is described below:

- In **Knowledge storing** step, beside $P_+^t(w|c)$ and $P_-^t(w|c)$, we also store $P_+^t(w_i|w_{i-1})$ and $P_-^t(w_i|w_{i-1})$ whereas $P_+^t(w_{i+1}|w_i) = \frac{\lambda+N_{+,w_iw_{i+1}}}{\lambda|V|+N_{+,w_i}}$ and $P_-^t(w_{i+1}|w_i) = \frac{\lambda+N_{-,w_iw_{i+1}}}{\lambda|V|+N_{-,w_i}}$). The number of occurrences of each bigram on each class ($N_{+,w_iw_{i+1}}^t$ and $N_{-,w_iw_{i+1}}^t$) and the domain-level knowledge ($M_{+,w_iw_{i+1}}^{KB}$, and $M_{-,w_iw_{i+1}}^{KB}$) are also stored.
- In **Optimization** step, due to the use of bigram, the probability for each document is modified as equations 3 4:

$$P(+|d) = \frac{P_+}{P_-}.P_+(w_0).P_+(w_1|w_0).P_+(w_2|w_1)\dots P_+(w_n|w_{n-1}) \quad (3)$$

$$P(-|d) = \frac{P_-}{P_+}.P_-(w_0).P_-(w_1|w_0).P_-(w_2|w_1)\dots P_-(w_n|w_{n-1}) \quad (4)$$

- The positive and negative probabilities for each document on the test data also have to follow the equations 3 4 for the **Sentiment Classification** step.

3.3 Bag of bigrams

Although using bigram help taking advantage of the phrases that express sentiments, using standard Bayes formula still relies on the probabilities and number of occurrences of unigrams on all the documents. Our alternative way to leverage bigram is to treat each bigram as a unigram and apply the normally used Bayes formula ($P_{+|d} = \frac{P_+}{P_-}.P_+(w_0w_1).P_+(w_1w_2).P_+(w_2w_3)\dots P_+(w_{n-1}w_n)$) to create the classifier. Such formula is applied to **Optimization** and **Sentiment classification** steps. We will compare how the two solutions improve the classification performance on both Vietnamese and English dataset.

3.4 Pre-Processing on Vietnamese dataset

Different to the dataset from Chen et al. [3] on English, the Tiki.vn dataset contains many emoticons. Therefore, we need to pre-process the data before **Knowledge storing** step to leverage all lexical resources in the dataset. In most online forums or discussion groups, users often use emoticons such as ":)", ":(" or punctuations such as '!!!!!' to express their opinions. However, during the task, we standardize the emoticons used by users, e.g. changing ":(((((" to ":(". We treat each emoticon or punctuation as a unigram and follow other steps as normal. In this pre-processing step, we also perform word segmentation by following the maximum entropy approach of Dinh and Vu [22]. Word segmentation can model the sentiment adjectives which often contain two or more morphemes, hence, provide a better vocabulary set for classification on Vietnamese using unigram feature.

4 Experimental Results

4.1 Dataset

In this study, we used two datasets for sentiment classification, one is Vietnamese and the other is English. The English one has been used by Chen et al. [3] for lifelong learning, in which there are 20,000 product reviews from Amazon divided into 20 domains. The Vietnamese dataset was also crawled from an e-commerce website, Tiki.vn, with consumer reviews on 17 products. The two datasets can contribute great values to different tasks of cross-domain sentiment analysis on both languages.

Labeled Vietnamese reviews For this study, we crawled the reviews from Tiki.vn, which is a large e-commerce website with quality reviews from the customers. It is a large corpus of 17 diverse domains or products and a total of 15,394 product reviews, which we name "A Community Resource for sentiment analysis on Vietnamese" (CRSAVi). We followed the previous work [23, 24] to treat reviews with more than 3 star as positive reviews, equal to 3 star as neutral and fewer than 3 star as negative ones. The number of positive, neutral and negative reviews are shown as in the table 1:

Table 1: Names of 17 domains and the number of positive, neutral and negative reviews

Product	**Positive**	**Neutral**	**Negative**
TrangDiem(Cosmetics)	3,629	792	154
Dungcuhocsinh(Tools for students)	1,803	164	37
Sanphamvegiay(Papers)	1,778	144	343
Butviet(Pens and pencils)	1,044	125	28
Dodungnhabep(Kitchen)	987	100	24
DauGoi(Shampoo)	347	59	18
Tainghe(Headphones)	698	90	18
DoDungChoBe(Baby)	658	61	14
Filehosobiahoso(Files)	157	47	14
Phukiendienthoaimaytin-hbang(Accessories)	583	32	13
Nuochoa(Perfume)	207	21	10
Thietbilamdep(Beauty equipment)	127	16	8
Butxoaxoakeo(Eraser)	311	43	7
Mayxaymayep(Grinders)	395	38	7
Binhdunsieutoc(Kettle)	107	13	5
Dungcuanuong(Dining substances)	114	19	5
TranhDongHo(Dong Ho paintings)	274	10	5
Total (15,394 reviews)	13,219	1,774	401

It is noted that the all product reviews from Tiki was checked by the website administrators before publishing, which helps guarantee the low rate of low quality reviews from online users. In fact, all of them contain Vietnamese tone marks,

some contain emoticons. On our dataset, the average unigram per document on each domain varies from 66 to just above 75 unigrams. The information packed in a single review in our dataset consists of product name, author name, rating, headline, bought-already, time of review and details. From the table 1, it can be seen that the proportion of negative class among the dataset is only around 2.6%. As that being said, to experiment life-long learning, a mass of reviews among multiple product types are required, although there is no Vietnamese sentiment dataset that can meet the requirements. Although different types of products are crawled for the task and Tiki has a great deal of book reviews, CRSAVi does not include books because most of the book reviews mention the book content, not the overall quality like other products.

Because the difference between the number of reviews across domains might result in the efficiency of the system, for this study, we selected a maximum amount of 100 reviews each class on each domain to conduct the experiments. From the total of 17 domains, to have a reasonable distribution of negative class on each set, we selected a group of 10 domains which have the most negative reviews. This group contains products which have equally and more than 13 negative reviews. This selection not only helped reduce the imbalanced distribution but also committed enough lexical resources for creating a classifier.

Labeled English reviews The corpus from Chen et al. [3] was utilized to compare directly with their life-long learning approach in English sentiment classification. The corpus contains reviews of 20 different products crawled from Amazon. The experiments were on a dataset which has a reasonable proportion of negative reviews across domains, varies from 11.97 to 30.51% .

4.2 Evaluation Metrics

The evaluation method used is 5-fold cross validation. While dividing a domain into groups, we tried to keep the class distribution to avoid the case of no negative review on a segment due to the small proportion of negative class mentioned above. F1-measure on negative and positive class in types of Micro-average and Macro-average are applied.

4.3 Baseline

Our method is compared to VietSentiWordnet by Vu et al [25]. The approach use a dictionary which contains a list of segmented words or phrases in Vietnamese that express sentiment. For each word or phrase, the dictionary provides correspondent positive and negative score. For each document, the score is evaluated by summing up all (positive score - negative score) of all sentiment words or phrases that are available in the dictionary. if the score is positive, the document is labeled as positive and vice versa. It is noted that VietSentiWordnet can only work on a single domain data.

On English, we compare our proposed method to Chen et al. [3] to illustrate the benefits of our approach on life-long learning.

4.4 Bigram feature improves the classification on English dataset

We compare our result to the original life-long learning approach of Chen et al. [3](LSC) on the balanced class distribution. We created a balance dataset of 200 reviews (100 positive and 100 negative) in each domain dataset for this experiment. On balanced class distribution, how the accuracy is improved is expressed as in table 2

Table 2: Accuracies on English balanced distribution over 20 domains

LSC	**LSC-bag-of-bigram**	**LSC-bigram**
83.34	**85.92**	85.44

Our method exceeds LSC to get to a high of 85.92%. This improvement confirms the results of Wang et al. [6] and prove that the use of bigram and bag-of-bigram features also improve the performance on cross-domain sentiment classification.

4.5 Vietnamese cross-domain sentiment classification

We compare our proposed method in different settings to the baseline method on the Vietnamese dataset. The average F1-score for the positive class is not shown because being the majority class makes the classifiers perform well and do not show much differences between multiple settings, although they all performs better than VietSentiWordnet. Table 3 compares VietSentiWordnet to our proposed method (life-long learning for Vietnamese sentiment classfication-now called LLVi) using unigram feature with segmentation (LLVi- uni) and without segmentation (LLVi-uniWS).

The LLVi with unigram feature (no segmentation) which also counts emoticons (LLVi-e) is also compared.

Table 3: Macro,micro average F1-score of the negative class on CRSAVi

VietSentiWordnet	**LLVi-uni**	**LLVi-e**	**LLVi-uniWS**
33.21,40.85	47.19,61.33	**51.20,62.12**	50.87,61.93

The table 3 has obviously shown that while segmentation task helps improving the performance on life-long learning with unigram feature. For example, the word “tuy_nhiên” (however) can classify well in our dataset but cannot be leveraged effectively without segmentation. However, life-long learning with emoticons still performs slightly better. The two emoticons “:(” and “:)” provides significantly biased probability thus become good classifiers. The table 3

also confirms that the life-long learning approach has a huge advantage over VietSentiWordnet, which can only works on the target domain.

Table 4 compares the performance is a collection of life-long learning approaches with different features applied. The group includes LLVi-uni, LLVi with bigram feature (LLVi-bi), LLVi with bag-of-bigram feature (LLVi-bb) in two settings, segmentation and without segmentation. Similar to English, using bigram

Table 4: Macro, micro average F1-score on negative class with Vietnamese dataset, unigram vs bigram vs bag-of-bigram. Unit: %

	LLVi-uni	**LLVi-bi**	**LLVi-bb**
no segmentation	47.19,61.33	56.10,56.52	60.56,**66.27**
with segmentation	50.87,61.93	64.37,61.53	**65.85**,60.59

and bag-of-bigram features make a huge improvement to the performance compared to using unigram feature only. With segmentation combined, the life-long learning using unigram feature improve significantly, while it does not have clear impact on life-long learning with bigram and bag-of-bigram features. 'cực_kỳ tốt'(extremely good), 'khó chịu'(frustrating), 'rát da'(burning skin sensation), 'chẳng mê'(cannot love) are examples of how using bigram performs better.

5 Conclusion

In this paper, we have presented our method that use life-long learning for cross-domain sentiment classification on English and Vietnamese. Experimental results on both corpus showed that:

- Life-long learning approach is effective for cross-domain sentiment classification in Vietnamese as well as in English.
- Incorporating bigram and bag-of-bigram features into life-long learning improved the performance of the system.
- Emoticons and word segmentation made slight improvement on sentiment classification on the Vietnamese dataset.

There is abundant room for further progress of our work. We would like to further exploit the sentiments from emoticons due to the high rate of occurrences of these in our dataset. Besides, future work could be focused on other collection of reviews with different qualities and different types of products to verify our proposed method.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis (2008)
2. Liu, B.: Sentiment analysis and opinion mining (2012)

3. Chen, Z., Ma, N., Liu, B.: Lifelong learning for sentiment classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, Association for Computational Linguistics (2015) 750–756
4. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. on Knowl. and Data Eng. **22** (2010) 1345–1359
5. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: Transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning. ICML '07, New York, NY, USA, ACM (2007) 759–766
6. Wang, S., Manning, C.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Korea, Association for Computational Linguistics (2012) 90–94
7. Chen, Z., Liu, B.: Mining topics in documents: standing on the shoulders of big data. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2014) 1116–1125
8. Chen, Z., Liu, B.: Topic modeling using topics from many domains, lifelong learning and big data. In Jebara, T., Xing, E.P., eds.: Proceedings of the 31st International Conference on Machine Learning (ICML-14), JMLR Workshop and Conference Proceedings (2014) 703–711
9. Ruvolo, P., Eaton, E.: Scalable lifelong learning with active task selection. In: AAAI Spring Symposium: Lifelong Machine Learning. (2013)
10. Finkel, J.R., Manning, C.D.: Hierarchical bayesian domain adaptation. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 602–610
11. Chen, M., Weinberger, K.Q., Blitzer, J.: Co-training for domain adaptation. In: Advances in neural information processing systems. (2011) 2456–2464
12. Kumar, A., Saha, A., Daume, H.: Co-regularization based semi-supervised domain adaptation. In: Advances in neural information processing systems. (2010) 478–486
13. Huang, F., Yates, A.: Exploring representation-learning approaches to domain adaptation. In: Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, Association for Computational Linguistics (2010) 23–30
14. Yang, H., Callan, J., Si, L.: Knowledge transfer and opinion detection in the trec 2006 blog track. In: TREC. (2006)
15. Blitzer, J., Dredze, M., Pereira, F., et al.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: ACL. Volume 7. (2007) 440–447
16. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th international conference on World wide web, ACM (2010) 751–760
17. Li, S., Zong, C.: Multi-domain sentiment classification. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics (2008) 257–260
18. Duyen, N.T., Bach, N.X., Phuong, T.M.: An empirical study on sentiment analysis for vietnamese. In: Advanced Technologies for Communications (ATC), 2014 International Conference on. (2014) 309–314

19. Bach, N.X., Phuong, T.M.: Leveraging user ratings for resource-poor sentiment classification. Procedia Computer Science **60** (2015) 322 – 331 Knowledge-Based and Intelligent Information Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.
20. Zhang, R., Tran, T.: An information gain-based approach for recommending useful product reviews. Knowledge and Information Systems **26** (2011) 419–434
21. Kieu, B.T., Pham, S.B.: Sentiment analysis for vietnamese. In: Knowledge and Systems Engineering (KSE), 2010 Second International Conference on, IEEE (2010) 152–157
22. Dien, D., Thuy, V.: A maximum entropy approach for vietnamese word segmentation. In: Research, Innovation and Vision for the Future, 2006 International Conference on. (2006) 248–253
23. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics (2007) 440–447
24. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2002) 79–86
25. Vu, X.S., Song, H.J., Park, S.B.: Building a vietnamese sentiwordnet using vietnamese electronic dictionary and string kernel. In: Knowledge Management and Acquisition for Smart Systems and Services. Springer (2014) 223–235