Journal of Computer Science and Cybernetics, V.34, N.4 (2018), 295–310
DOI 10.15625/1813-9663/34/4/13160

VLSP SHARED TASK: SENTIMENT ANALYSIS

NGUYEN THI MINH HUYEN[1,*], NGUYEN VIET HUNG[1], NGO THE QUYEN[1], VU XUAN LUONG[2], TRAN MAI VU[3], NGO XUAN BACH[4], LE ANH CUONG[5]

[1] *VNU University of Science;* [2] *Vietlex*
[3] *VNU University of Engineering and Technology*
[4] *Post and Telecommunication Institute of Technology*
[5] *Ton Duc Thang University*
[*] *huyenntm@hus.edu.vn*

Abstract. Sentiment analysis is a Natural Language Processing (NLP) task of identifying or extracting the sentiment content of a text unit. This task has become an active research topic since the early 2000s. During the two last editions of the VLSP workshop series, the shared task on Sentiment Analysis (SA) for Vietnamese has been organized in order to provide an objective evaluation measurement about the performance (quality) of sentiment analysis tools, and encourage the development of Vietnamese sentiment analysis systems, as well as to provide benchmark datasets for this task. The first campaign in 2016 only focused on the sentiment polarity classification, with a dataset containing reviews of electronic products. The second campaign in 2018 addressed the problem of Aspect Based Sentiment Analysis (ABSA) for Vietnamese, by providing two datasets containing reviews in restaurant and hotel domains. These data are accessible for research purpose via the VLSP website `vlsp.org.vn/resources`. This paper describes the built datasets as well as the evaluation results of the systems participating to these campaigns.

Keywords. Aspect based sentiment analysis; Evaluation, opinion mining; Sentiment analysis; Shared task, Vietnamese, VLSP workshop.

1. INTRODUCTION

With the development of technology and the Internet, different types of social media such as social networks and forums have allowed people to not only share information but also to express their opinions and attitudes on products, services and other social issues. The Internet becomes a very valuable and important source of information. People nowadays use it as a reference to make their decisions on buying a product or using a service. Moreover, this kind of information also lets the manufacturers and service providers receive feedback about the limitations of their products and therefore should improve them to meet the customer needs better. Furthermore, it can also help authorities know the attitudes and opinions of their residents on social events so that they can make appropriate adjustments.

Since the early 2000s, opinion mining and sentiment analysis [3] have become a new and active research topic in natural language processing and data mining. The major tasks in this topic include:

- Subjective classification: This is the task of detecting whether a document contains personal opinions or not (only provides facts).
- Polarity classification (Sentiment classification): Classify the opinion expressed in a document into one of three types, which are "positive", "negative" and "neutral".
- Spam detection: Detect fake reviews and reviewers.
- Rating: Reflect the personal opinion expressed in a document as a rating from 1 star to 5 stars (very negative to very positive).
- Opinion summarization: Generate effective summaries of opinions so that users can get a quick understanding of the underlying sentiments.

Besides these basic tasks, there are deeper studying tasks as follows:

- Aspect-based sentiment analysis (ABSA): The goal is to identify the aspects of given target entities and the sentiment expressed for each aspect.
- Opinion mining in comparative sentences: This task focuses on mining opinions from comparative sentences, i.e., to identify entities to be compared and determine which entities are preferred by the author in a comparative sentence.

For popular language such as English, there are many campaigns for this research topic. The international workshop series on Semantic Evaluation (SemEval) has organized successfully such campaigns for several years, as described in [4] (polarity classification) and [1] (ABSA).

Meanwhile, for Vietnamese language, until 2016 there is no systematic comparison between the performance of Vietnamese sentiment analysis systems. The first related campaign for Vietnamese language sentiment analysis was organized at VLSP 2016 (SA-VLSP2016), which only focused on polarity classification. This benchmark dataset contained short reviews on technical articles from forums and social networks, with polarity annotation (*positive*, *negative* and *neutral*). The second campaign organized in the framework of the VLSP 2018 workshop addresses the problem of ABSA for Vietnamese (ABSA-VLSP2018), in which we provide two datasets containing reviews in restaurant and hotel domains annotated with aspects and the corresponding sentiment polarities. These benchmark datasets are accessible for research purpose via the VLSP website `vlsp.org.vn/resources`.

The remainder of this report is organized as follows. First, we describe the shared tasks, the dataset construction and the evaluation measures. Then we summarize and discuss about the participating systems and their results and finally we make some conclusions on these campaigns.

2. TASK DESCRIPTION

2.1. SA-VLSP2016

2.1.1. Task definition

The scope of this first campaign is polarity classification, i.e., to evaluate the ability of classifying Vietnamese reviews/documents into one of three categories: positive, negative, or neutral. The data domain is technical article reviews.

Logitech pin trâu thôi rồi, mua 1 con B175 cùi mà cục pin theo chuột 3 năm chưa phải thay! ai chê thì chê chứ tôi thấy chuột Logitech xài hơi bị thích !

POS

Figure 1. Example of input review and expected output

Table 1. SA-VLSP 2016: Quantities of comments from three data sources

No.	Source	Quantity
1	tinhte.vn	2710
2	vnexpress.net	7998
3	facebook	1488
	Total	**12190**

Figure 1 shows an example from the training dataset.

2.1.2. Data collection

The data were collected from three source sites which are `tinhte.vn`, `vnexpress.net` and Facebook. Our data consists of comments of technical articles on those sites. The quantities of comments are reported in Table 1.

2.1.3. Annotation procedure

We have three annotators for our dataset. First, we split 12196 comments into three parts, one for each annotator. Each annotator had to give each comment one of four labels which are POS (positive), NEG (negative), NEU (neutral) and USELESS. Because a review can be very complex with different sentiments on various objects, we set some constraints on the dataset and used USELESS label to filter out the irrelevant comments. The constrains are:

- The dataset only contains reviews having personal opinions.
- The data are usually short comments, containing opinions on one object. There is no limitation on the number of the objects aspects mentioned in the comment.
- Label (POS/NEG/NEU) is the overall sentiment of the whole review.
- The dataset contains only real data collected from social media, not artificially created by human.

Normally, it is very difficult to rate a neutral comment because the opinions are always indeclinable to be negative or positive.

- We usually rate a review be neutral when we cannot decide whether it is positive or negative.
- The neutral label can be used for the situations in which a review contains both positive and negative opinions but when combining them, the comment becomes neutral.

After filtering the data, we had 2669 POS, 2359 NEG and 2122 NEU. Next, we changed the annotator for each part. After the annotators had labeled the their parts, we selected 2100 comments in each part for the next step. In the next step, we changed the annotator for each part again. The result of this step was compared to the ones in two previous steps. Then, discussions were made in order to reach agreement to the final result. The last step is selecting data for the evaluation campaign by removing all divergent comments (different labels by two annotators, including the data discussed and reached agreement). Finally, for each label, we had 1700 comments for training, 350 comments for testing.

2.1.4. Evaluation measures

The performance of the sentiment classification systems are evaluated using accuracy, precision, recall, and the F1 score.

$$\text{accuracy} = \frac{\text{number of correctly classified reviews}}{\text{number of reviews}}. \tag{1}$$

Let A and B be the set of reviews that the system predicted as POS and the set of reviews with POS label in the gold data, the precision, recall, and the F1 score of POS label can be computed as follows (similarly for NEG label):

$$\text{Precision} = \frac{|\, A \cap B \,|}{|\, A \,|}, \tag{2}$$

$$\text{Recall} = \frac{|\, A \cap B \,|}{|\, B \,|}, \tag{3}$$

$$POS\_F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

$$Average\_F1 = \frac{POS\_F1 + NEG\_F1}{2}. \tag{5}$$

2.2. ABSA-VLSP 2018

2.2.1. Task definition

The second campaign for Vietnamse sentiment analysis covers a more complicated problem: the aspect-based sentiment analysis. This task is similar to the Subtask 2 (slot 1 and slot 3) of the SemEval 2016 Task 5 [1]. Given a customer review about a target entity, the goal is to identify a set of {*aspect*, *polarity*} tuples that summarize the opinions expressed in this review. Aspect is a pair of *entity-attribute*, while polarity can be "*positive*", "*negative*" or "*neutral*".

The task considers reviews in two domains: Restaurant and Hotel. Figure 2 shows two examples of input reviews in the two domains and expected outputs. In Example 1, the goal is to recognize the following three tuples:

Example 1: Restaurant domain

#lozi #lozisaigon #chaovit #anchinh ngon - bổ - rẻ. Khuyết là hơi xa trung tâm (delicious - good for health - cheap. Far form center)

{FOOD#PRICE, positive}, {FOOD#QUALITY, positive}, {LOCATION#GENARAL, negative}

Example 2: Hotel domain

Phòng ốc sạch, giường thoải mái, nhân viên thân thiện. (clean rooms, comfortable beds, friendly staffs)

{ROOMS#CLEANLINESS, positive}, {ROOMS#COMFORT, positive}, {SERVICE#GENERAL, positive}

Figure 2. Examples of input reviews and expected outputs

1. {*aspect* = FOOD#PRICE, *polarity* = positive};
2. {*aspect* = FOOD#QUALITY, *polarity* = positive};
3. {*aspect* = LOCATION#GENERAL, *polarity* = positive}.

Similarly, in Example 2, we aim to extract the following three tuples:

1. {*aspect* = ROOMS#CLEANLINESS, *polarity* = positive};
2. {*aspect* = ROOMS#COMFORT, *polarity* = positive};
3. {*aspect* = SERVICE#GENERAL, *polarity* = positive}.

The task is divided into two subtasks (two phases):

- **Phase A (Aspect)**: The participants are required to identify aspects (entity - attribute) only.
- **Phase B (Aspect - Polarity)**: The participants are required to identify both aspects and sentiment polarities.

2.2.2. Data collection

Raw data were crawled from:

- `https://lozi.vn/` (for restaurant domain).
- `https://www.booking.com/` (for hotel domain).

We selected reviews from hotels in Ha Noi, Da Nang, and Ho Chi Minh City (150 hotels in each city) to annotate manually. The labeled dataset contains 4751 reviews for restaurant domain and 5600 reviews for hotel domain.

2.2.3. Annotation procedure

Data were annotated by three people. For each domain, we divided the dataset into two subsets. First, two annotators were asked to identify aspects and polarities in two subsets (each annotator for one subset). Then, the third annotator checked labeled data. If annotators disagreed on an assignment, three people were asked to examine and make the final decision.

In the following, we describe the set of aspects for each domain.

- **Aspects for restaurant domain**: Entities can be RESTAURANT (in general), AMBIENCE, LOCATION, FOOD, DRINKS, or SERVICE; attributes can be GENERAL, QUALITY, PRICE, STYLE & OPTIONS, or MISCELLANEOUS. The possible combinations of these entities and attributes are given in Table 2. Totally, we have 12 aspect categories for restaurant domain.
- **Aspects for hotel domain**: Entities can be HOTEL (in general), ROOMS, ROOM AMENITIES, FACILITIES, SERVICE, LOCATION, or FOOD & DRINKS; attributes can be GENERAL, PRICES, DESIGN & FEATURES, CLEANLINESS, COMFORT, QUALITY, STYLE & OPTIONS, or MISCELLANEOUS. The possible combinations of these entities and attributes are given in Table 3. Totally, we have 34 aspect categories for hotel domain.

Table 2. Possible entity-attribute pairs for restaurant domain

	GENERAL	**PRICES**	**QUALITY**	**STYLE & OPTIONS**	**MISCELLANEOUS**
RESTAURANT	√	√	×	×	√
FOOD	×	√	√	√	×
DRINKS	×	√	√	√	×
AMBIENCE	√	×	×	×	×
SERVICE	√	×	×	×	×
LOCATION	√	×	×	×	×

For each domain, data were divided into three datasets: training, development, and test. Training and development datasets were used to train participating systems. Test dataset was used for the final evaluation purpose. Table 4 shows the number of reviews and aspects in each dataset.

2.2.4. Evaluation measure

The performance of participating systems were evaluated in two phases.

Table 3. Possible entity-attribute pairs for hotel domain.

	GENERAL	**PRICES**	**DESIGN & FEATURES**	**STYLE & CLEANLI-NESS**	**COMFORT**	**QUALITY**	**STYLE & OPTION**	**MISCELLANE-OUS**
HOTEL	√	√	√	√	√	√	×	√
ROOMS	√	√	√	√	√	√	×	√
ROOM_AMENTITIES	√	√	√	√	√	√	×	√
FACILITIES	√	√	√	√	√	√	×	√
SERVICE	√	×	×	×	×	×	×	×
LOCATION	√	×	×	×	×	×	×	×
FOOD & DRINK	×	√	×	×	×	√	√	√

Table 4. Statistical information of training, development, and test datasets

Domain	**Dataset**	**#Reviews**	**#Aspects**
Restaurant	Training	2961	9034
	Development	1290	3408
	Test	500	2419
Hotel	Training	3000	13948
	Development	2000	7111
	Test	600	2584

- **Phase A**: Aspect (Entity-Attribute).

  The F1 score will be calculated for aspects only. Let A be the set of predicted aspects (entity-attribute pairs), and B be the set of annotated aspects, precision, recall, and the F1 score are computed as follows:

$$\text{Precision} = \frac{|\, A \cap B \,|}{|\, A \,|}, \tag{6}$$

$$\text{Recall} = \frac{|\, A \cap B \,|}{|\, B \,|}, \tag{7}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

- **Phase B**: Full (Aspect-Polarity).

  The F1 score will be calculated for both aspects and sentiment polarities. Let A be the set of predicted tuples (entity-attribute-polarity), and B be the set of annotated

tuples, the precision, recall, and the F1 score can be computed in a similar way as in Phase A.

3. SUBMISSIONS AND RESULTS

3.1. Submissions in SA-VLSP2016

There are eight teams participating in this campaign. We received full reports from five teams and short descriptions from two teams. The last one did not send us any report. Generally, all of the participating systems treat our task as a classification problem and use statistical machine learning approaches with various feature extraction and selection techniques to solve it. From the experiments of the systems, we have some interesting points to discuss in the next sections.

3.1.1. Methods and Features

The methods used by participating systems are presented in Table 5. Support Vector Machine (SVM) is the most popular method chosen by the teams. Besides, neural network architectures such as multilayer neural network (MLNN) and long short-term memory (LSTM) network, are also used by two teams due to its success in the recent years. Other methods are maximum entropy (MaxEnt), perceptron, random forest, naive Bayes and gradient boosting which have been proved to be useful in NLP tasks. While almost all teams tended to do experiments in individual models, there is one team (sa3) which tried to combine three models into one system using an ensemble methods [6].

In term of features, almost all systems use the basic n-gram features. TF-IDF also plays an important role in many systems [6], [8], [2]. In addition, some systems use external dictionaries of sentiment words, booster words, reversed words and emotion words to enrich their feature sets and help to gain better results [10], [7].

3.1.2. Results

The best results of all teams are reported in Table 6 where systems are ranked by their average F1 scores. In case that a team had more than one system, the best one is marked with "best" in Table 5. The highest score belongs to sa1 team [7] who used MaxEnt model with n-gram features and phrase features extracted from hand-built dictionaries. In [7], the authors reported that with the same feature set, MaxEnt model significantly outperforms SVM by a gap of approximately 7% in terms of F1 score. This strongly surprised us. The result of sa1 is also much better than others'. We are aware that their hand-built dictionaries of sentiment and intensity words may have an important effect on the result of the system in our test set.

The team sa2 [2] only uses TF-IDF features in an MLNN to achieve a promising result 71.44% for average F1. They also have experiments on SVM and LSTM with features extracted from VietSentiWordNet but the results are not as good as MLNNs. The ensemble system of sa3 [6] combines three sub-systems which are random forest, SVM and naive Bayes. This system produces a good result at 71.22% for F1 score. The ensemble system also uses only TF-IDF weighted n-gram features. Team sa4 [10] used SVM as learning method combining with n-gram features and various other features extracted from external

Table 5. Methods of VLSP 2016 participating systems

Team	Methods	Features
sa1	Perceptron SVM MaxEnt (best)	n-gram (1, 2, 3) on syllables, dictionary of sentiment words and phrases
sa2	SVM MLNN (best) LSTM	TF-IDF on 1,2-gram (best) VietSentiWordNet TFIDF-VietSentiWordNe
sa3	Random forest SVM Naive Bayes	TF-IDF weighted n-gram (1, 2, 3)
sa4	SVM	n-gram booster word list, reverser word list, emotion word list
sa5	SVM MLNN (best)	BOW, TF-IDF (best) BOW-senti, TF-IDF-senti, Objectivity-score
sa6	SVM	n-gram (1, 2 ,3) extracted on words, sybllables and important words. Word embedding (using GloVe) Log-count ratio of n-gram, Negation words
sa7	Gradient boosting	TF-IDF on words (remove words having low TF-IDF)
sa8	No report	No report

Table 6. Results of systems participating to SA shared task at VLSP 2016

Team	Positive			Negative			Average F1
	P	R	F1	P	R	F1	
sa1	75.85	89.71	82.2	79.88	76	77.89	80.05
sa2	72.42	74.29	73.34	69.94	69.14	69.54	71.44
sa3	74.77	71.14	72.91	72.09	67.14	69.53	71.22
sa4	68.11	72	70	60.59	70.29	65.08	67.54
sa5	69.06	71.43	70.23	65.67	62.86	64.23	67.23
sa6	71.8	70.57	71.18	67.1	59.43	63.03	67.11
sa7	71	67.14	69.02	62.97	61.71	62.33	65.68
sa8	21.25	4.86	7.91	44.72	67.71	53.86	30.89

dictionaries that help to gain average F1 score at 67.54%. Next, the report of team sa5 [8] also shows that MLNN outperforms SVM in our task. Various features is used by their system and they also found that TF-IDF helps to gain the best result. Meanwhile, the SVM-based system of team sa6 uses various kind of features including n-gram on words, syllables, important words such as verb, noun, adjective, etc., word embedding, etc., however, its result is not as good as other SVM-based systems that make use of TF-IDF features.

3.2. Submissions in ABSA-VLSP2018

At VLSP 2018, 13 teams have registered and got the training and development datasets for the ABSA shared task. However, we finally only received submissions from 3 teams. Among them, two teams submitted technical reports and the other one sent us a short description. All teams considered the task as classification problems and exploited statistical machine learning algorithms to solve. In the next section, we summarize methods and results of 3 participating systems: SA1 from Van et al. [9], SA2 from Nguyen and Minh [5], and SA3 from Vu and Anh.

3.2.1. Methods

While SA2 and SA3 considered the task as a multi-class classification problem (each label is a pair of *aspect-polarity*) and built only one classifier to solve the task, SA1 treated the task as multiple binary classification problems and built a single binary classifier for each aspect. To identify polarities of reviews, SA1 modeled the problem as a classification with three classes, i.e. positive, negative, and neutral.

Table 7 summarizes learning algorithms and features used in participating systems. While SA1 and SA3 used SVM with linear kernel, SA2 exploited multilayer perceptron algorithm. SA2 and SA3 built only one multi-class classifier with basic features, including n-grams and TF-IDF scores. SA1 used more sophisticated features, such as elongate features, hagtags, punctuation marks. SA1 also conducted some preprocessing steps before training classification models.

Table 7. Learning algorithms and features used in VLSP 2018 participating systems

System	Learning Algorithms	Features
SA1	Linear SVM (sklearn-toolkit)	**Aspect**: n-grams, words, POS tags **Polarity**: n-grams, words, Elongate, Aspect Category, Count of the hagtags, Count of POS tags, Punctuation Marks
SA2	Multilayer Perceptron (scikit-learn library)	n-grams, TF-IDF
SA3	Linear SVM	Count features (n-grams), TF-IDF

3.2.2. Results

Tables 8 and 9 summarize results of participating systems on development and test datasets, respectively. For both domains, SA1 achieved the best F1 scores on both development and test datasets. The results showed the effectiveness of sophisticated features used in SA1. Using linear SVM, SA1 and SA3 outperformed SA2 with multilayer perceptron significantly.

The detailed results of the teams on each aspect are shown in charts. Aspects and acronyms are shown in the Table 3.2.2. and Table 3.2.2. for Hotel and Restaurant data. The amount of data on each aspect in test data is presented in Figure 3 and 4.

Table 8. Results on development datasets of VLSP 2018 participating systems

		Phase A (Aspect)			**Phase B (Aspect-Polarity)**		
Domain	**Team**	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
Restaurant	SA1	0.75	0.85	**0.79**	0.63	0.71	**0.67**
	SA2						0.59
	SA3	0.78	0.65	0.71	0.71	0.59	0.64
Hotel	SA1	0.75	0.64	**0.69**	0.67	0.58	**0.62**
	SA2						0.56
	SA3	0.83	0.51	0.63	0.78	0.48	0.6

Table 9. Results on test datasets of VLSP 2018 participating systems.

		Phase A (Aspect)			**Phase B (Aspect-Polarity)**		
Domain	**Team**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Restaurant	SA1	0.79	0.76	**0.77**	0.62	0.6	**0.61**
	SA2	0.88	0.38	0.54	0.79	0.35	0.48
	SA3	0.62	0.62	0.62	0.52	0.52	0.52
Hotel	SA1	0.76	0.66	**0.7**	0.66	0.57	**0.61**
	SA2	0.85	0.42	0.56	0.8	0.39	0.53
	SA3	0.83	0.58	0.68	0.71	0.49	0.58

Table 10. The aspects in Hotel data

Acronym	**Aspect**	**Acronym**	**Aspect**
asp#1	ROOM_AMENITIES#CLEANLINESS	**asp#18**	HOTEL#PRICES
asp#2	SERVICE#GENERAL	**asp#19**	HOTEL#GENERAL
asp#3	ROOMS#CLEANLINESS	**asp#20**	ROOMS#PRICES
asp#4	ROOMS#COMFORT	**asp#21**	HOTEL#COMFORT
asp#5	LOCATION#GENERAL	**asp#22**	FACILITIES#GENERAL
asp#6	ROOMS#GENERAL	**asp#23**	HOTEL#MISCELLANEOUS
asp#7	ROOMS#DESIGN&FEATURES	**asp#24**	ROOM_AMENITIES#QUALITY
asp#8	HOTEL#CLEANLINESS	**asp#25**	FACILITIES#MISCELLANEOUS
asp#9	ROOM_AMENITIES#COMFORT	**asp#26**	FACILITIES#COMFORT
asp#10	ROOM_AMENITIES#DESIGN&FEATURES	**asp#27**	FOOD&DRINKS#QUALITY
asp#11	ROOM_AMENITIES#GENERAL	**asp#28**	FOOD&DRINKS#MISCELLANEOUS
asp#12	FOOD&DRINKS#STYLE&OPTIONS	**asp#29**	FACILITIES#PRICES
asp#13	ROOMS#QUALITY	**asp#30**	FOOD&DRINKS#PRICES
asp#14	FACILITIES#DESIGN&FEATURES	**asp#31**	FACILITIES#CLEANLINESS
asp#15	HOTEL#DESIGN&FEATURES	**asp#32**	ROOM_AMENITIES#MISCELLANEOUS
asp#16	FACILITIES#QUALITY	**asp#33**	ROOM#MISCELLANEOUS
asp#17	HOTEL#QUALITY	**asp#34**	ROOM_AMENITIES#PRICES

The result of SA1 is presented in Figure 5 and Figure 6. With Hotel data, SA1 achieved the highest results on asp#2, asp#5 and asp#3 as 0.85, 0.83 and 0.73. The top 3 highest

Table 11. The aspects in Restaurant data

Acronym	Aspect	Acronym	Aspect
asp#1	RESTAURANT#GENERAL	**asp#7**	FOOD#STYLE&OPTIONS
asp#2	DRINKS#QUALITY	**asp#8**	FOOD#PRICES
asp#3	DRINKS#PRICES	**asp#9**	FOOD#QUALITY
asp#4	DRINKS#STYLE&OPTIONS	**asp#10**	AMBIENCE#GENERAL
asp#5	LOCATION#GENERAL	**asp#11**	RESTAURANT#PRICES
asp#6	RESTAURANT#MISCELLANEOUS	**asp#12**	SERVICE#GENERAL

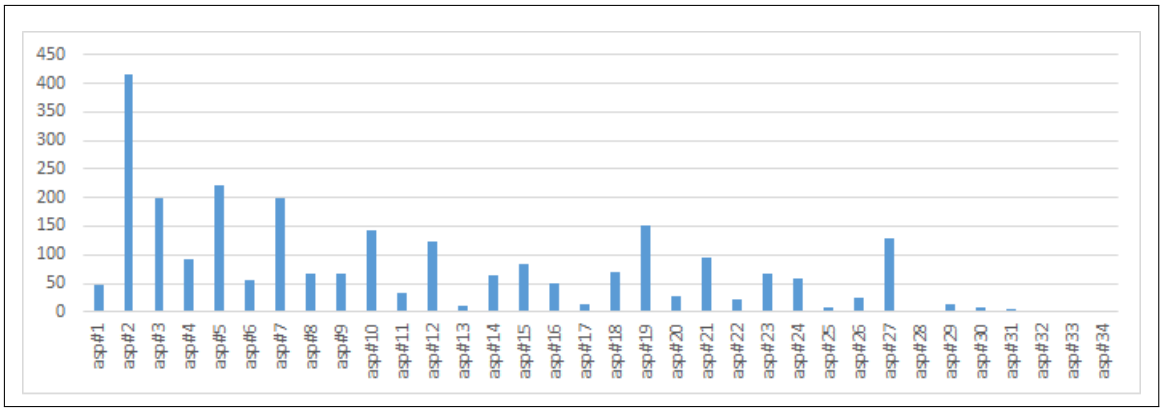

Figure 3. The chart present the amount of data for each aspect in Hotel

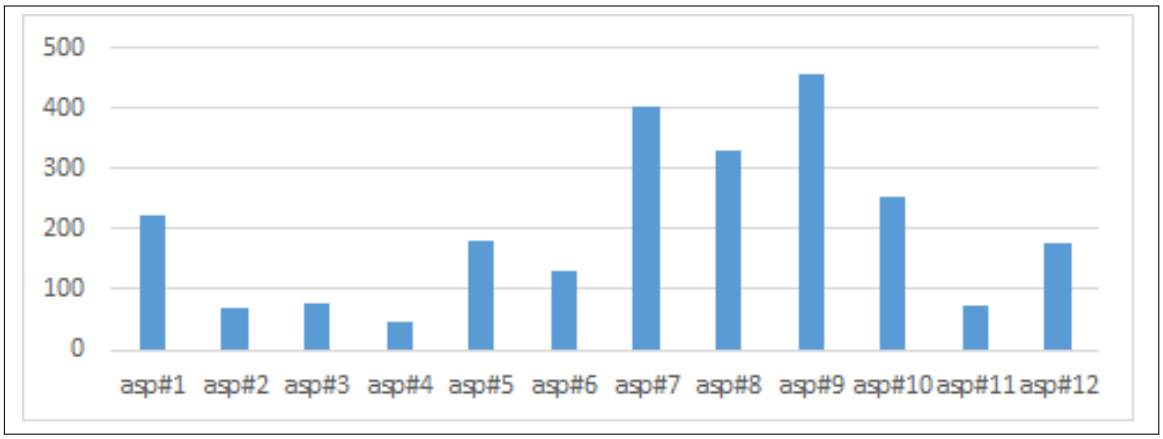

Figure 4. The chart present the amount of data for each aspect in Restaurant

rated results for the Restaurant data are 0.85, 0.76 and 0.6 on asp#9, asp#7 and asp#1. This is the best team in the competition this year.

The result of SA2 is presented in Figure 7 and Figure 8. With Hotel data, SA2 achieved the highest results on asp#2, asp#5 and asp#19 as 0.82, 0.76 and 0.64. The top 2 highest rated results for the Restaurant data are 0.85 and 0.66 on asp#9, asp#7, but results on

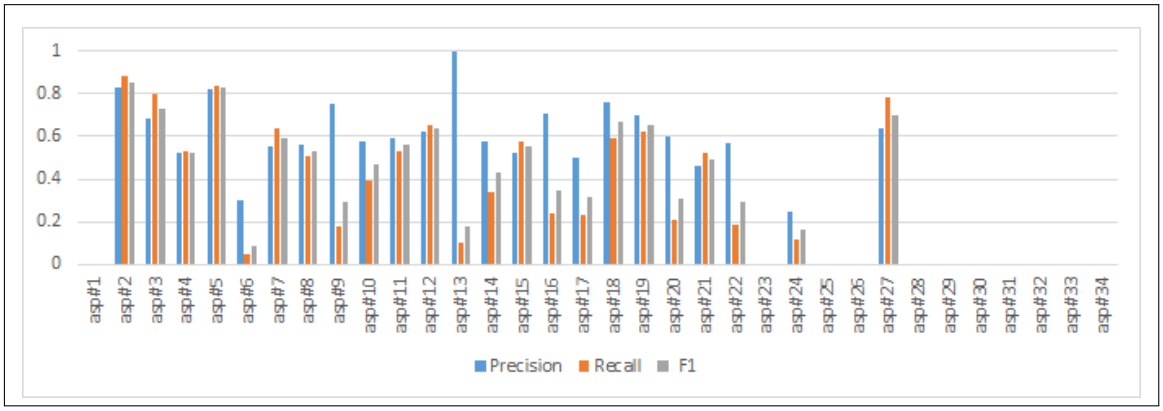

Figure 5. The chart present the result of SA1 on Hotel data

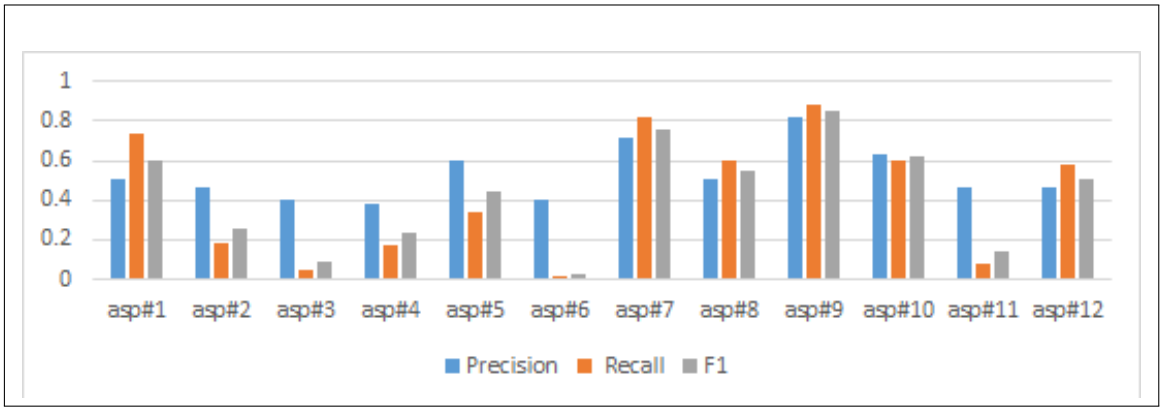

Figure 6. The chart present the result of SA1 on Restaurant data

other aspects are low, under 0.4.

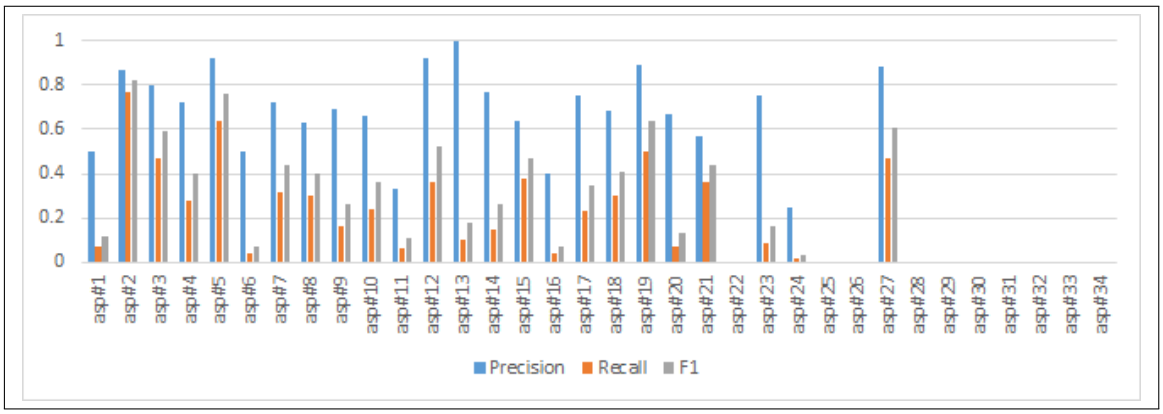

Figure 7. The chart present the result of SA2 on Hotel data

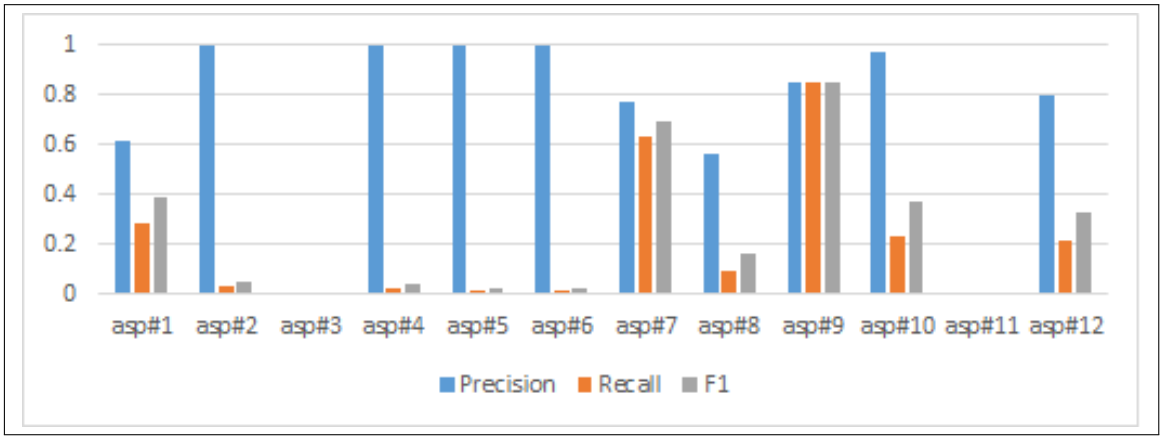

Figure 8. The chart present the result of SA2 on Restaurant data

The result of SA3 is presented in Figure 9 and Figure 10. With Hotel data, SA3 achieved the highest results on asp#2, asp#5 and asp#3 as 0.74, 0.71 and 0.67. The top 3 highest rated results for the Restaurant data are 0.85, 0.75 and 0.66 on asp#9, asp#7 and asp#10.

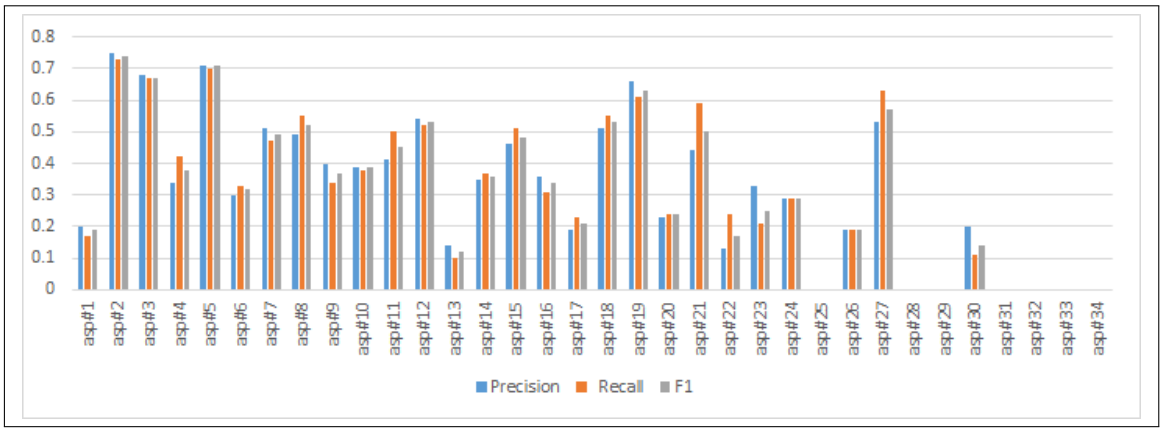

Figure 9. The chart present the result of SA3 on Hotel data

Based on the results of the teams, we found that all the teams achieved better results on the same aspects. This may be due to the amount of data of these aspects more than other aspects, and these aspects are less ambiguous than other.

4. CONCLUSION

In this paper, we have described the results of the shared tasks on Sentiment Analysis, organized in the framework of two last editions of VLSP workshop series: VLSP 2016 and VLSP 2018. These two campaigns have attracted an important number of research teams as well as the public attention.

Three benchmark datasets for Vietnamese language have been built and made available

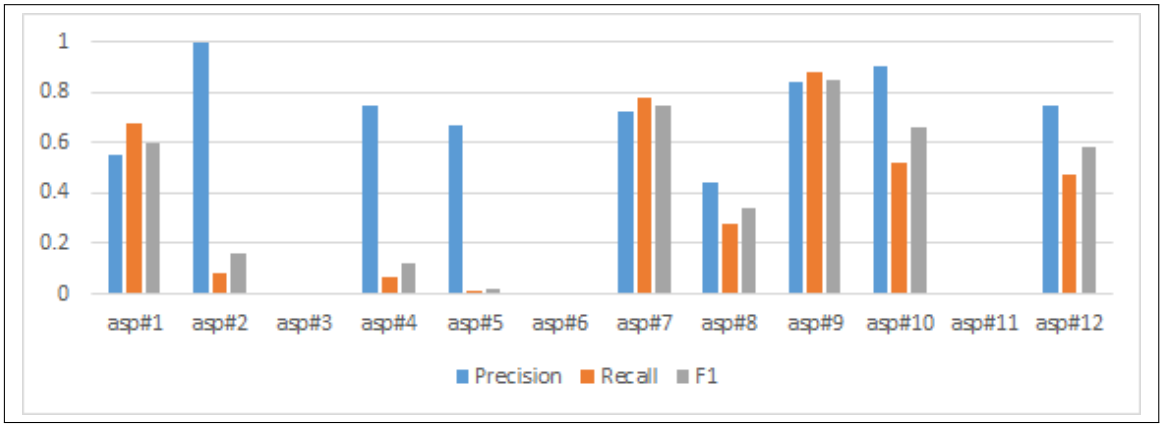

Figure 10. The chart present the result of SA3 on Restaurant data

for research purpose in the field of sentiment analysis: one for the task of polarity classification, two other datasets for a deeper task which is aspect based sentiment analysis. These datasets remain quite simple, as the manual annotation covers only the expected output but not other linguistic annotation. However, we strongly believe that these resources will help to impulse the development of researching on this topic in the near future.

The first campaign in 2016 had a good number of participants: 8 teams. The second evaluation campaign in 2018 had only 3 finalists, which is much smaller than the number of participants in the previous VLSP workshop. The reason might be that the task this year is more difficult than the previous one. All participating systems implemented popular machine learning approach and used many rich features to solve the task.

In the next steps, we continue to scale the size of the datasets as well as to enrich the linguistic annotation of these resources for sentiment analysis tasks. We hope to receive more attention from the research community and companies in the next VLSP workshops.

ACKNOWLEDGMENT

The organization of these shared tasks was partially supported by the following sponsors: Alt Vietnam, InfoRe, VCCorp, Viettel Cyberspace Center and Zalo Careers. Great thanks to them! We would like to address our special thanks to Dr. Hoang Thi Tuyen Linh and Mr. Vu Hoang, as well as our students from Vietnam National University, Hanoi, for their contribution to the data annotation process. And we thank to all the research teams for their participation to this competition.

REFERENCES

[1] P. et al., "Semeval-2016 task 5: Aspect based sentiment analysis." in *Proceedings of SemEval-2016*, 2016, pp. 19–30.

[2] N. Hy, L. Tung, L. Viet-Thang, and D. Dien, "A simple supervised learning approach to sentiment classification at VLSP 2016," in *The Fourth International Workshop on*

Vietnamese Language and Speech Processing (VLSP 2016), 2016. [Online]. Available: http://vlsp.org.vn/archives

[3] B. Liu, *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, 2012.

[4] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US, 2016, pp. 1–18.

[5] T. A. Nguyen and P. Q. N. Minh, "Using multilayer perceptron for aspect-based sentiment analysis at VLSP-2018 sa task," in *Proceedings of the Fifth International Workshop on Vietnamese Language and Speech Processing (VLSP 2018)*, 2018. [Online]. Available: http://vlsp.org.vn/archives

[6] Q. N. M. Pham and T. T. Tran, "A lightweight ensemble method for sentiment classification task," in *The Fourth International Workshop on Vietnamese Language and Speech Processing (VLSP 2016)*, 2016. [Online]. Available: http://vlsp.org.vn/archives

[7] T. P. Quynh-Trang, N. Xuan-Truong, T. Van-Hien, N. Thi-Cham, and T. Mai-Vu, "Dsktlab: Vietnamese sentiment analysis for product reviews," in *The Fourth International Workshop on Vietnamese Language and Speech Processing (VLSP 2016)*, 2016. [Online]. Available: http://vlsp.org.vn/archives

[8] T. T. Tran, X. Ho, and N. T. Nguyen, "A multi-layer neural networkbased system for vietnamese sentiment analysis at the VLSP 2016 evaluation campaign," in *The Fourth International Workshop on Vietnamese Language and Speech Processing (VLSP 2016)*, 2016. [Online]. Available: http://vlsp.org.vn/archives

[9] T. D. Van, K. V. Nguyen, and N. L.-T. Nguyen., "Nlp@uit at VLSP 2018: A supervised method for aspect based sentiment analysis," in *Proceedings of the Fifth International Workshop on Vietnamese Language and Speech Processing (VLSP 2018)*, 2018. [Online]. Available: http://vlsp.org.vn/archives

[10] N. V. Vi, H. V. Minh, and N. T. Tam, "Sentiment analysis for vietnamese using support vector machines with application to facebook," in *The Fourth International Workshop on Vietnamese Language and Speech Processing (VLSP 2016)*, 2016. [Online]. Available: http://vlsp.org.vn/archives

Received on October 03, 2018
Revised on December 04, 2018