VLSP 2016 Shared Task: Sentiment Analysis

Le Anh Cuong
Faculty of Information Technology
Ton Duc Thang University, Ho Chi Minh city, Vietnam

Nguyen Thi Minh Huyen and Nguyen Viet Hung
Faculty of Mathematics, Mechanics and Informatics
VNU University of Science, Hanoi, Vietnam

***Abstract*—The Fourth International Workshop on Vietnamese Language and Speech Processing (VLSP 2016) organizes the shared task of Sentiment Analysis at the first time for Vietnamese language processing. This campaign aims to bring together researchers interested in this topic as well as provides a benchmark dataset for this task. We received eight submissions with various methods and promising results.**

I. 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 early 2000s, opinion mining and sentiment analysis have become a new and hot research topic in Natural language Processing and Data Mining. The paper [1] is a very good survey for the development of this topic. The major tasks in this topic include:

- *Subjective classification*: aims to classify subjectivity and objectivity documents.
- *Polarity sentiment classification*: aims to classify an subjectivity document into one of the three classes: “positive”, “negative” and “neutral”.
- *Spam detection*: aims to detect fake reviews and reviewers.
- *Rating*: rating the documents having personal opinions from 1 star to 5 star (very negative to very positive).

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

- *Aspect-based sentiment analysis*: 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 campaign for this research topic. One of the most successful campaigns is described in [2]. Meanwhile, for Vietnamese language, so far there is no systematic comparison between the performance of Vietnamese sentiment analysis systems. The VLSP 2016 campaign, therefore, targets at providing an objective evaluation measurement about performance (quality) of sentiment analysis tools, and encouraging the development of Vietnamese sentiment analysis systems. As the first shared task on Sentiment Analysis, we just focus on the essential problem, that is polarity sentiment classification. We created a standard corpus for evaluating Vietnamese sentiment analysis systems which contains comments on technical articles from forums and social networks. This is actually the first benchmark dataset for this task. There are total eight submissions to the workshop and almost of them produce very promising results.

The remainder of this report is organized as follows: first, we describe the task, the preparation of the dataset and evaluation method; then, we summarize and discuss about the participating systems and their results and finally we make some conclusions on this campaign.

II. Task Description

A. Problem

The scope of the campaign this year is polarity classification, i.e., to evaluate the ability of classifying Vietnamese reviews/documents into one of three categories: positive, negative, or neutral. Other sentiment analysis tasks can be covered in the campaigns next years.

B. Data preparation

1) Data collection: We collected data from three source sites which are Tinhte.vn, Vnexpress.net and Facebook. Our data consists of comments of technical articles in those sites. The quantities of comments are reported in Table I.

TABLE I
Analysis of data source

No.	Source	Quantity
1	Tinhte.vn	2710
2	Vnexpress.net	7998
3	Facebook	1488
	Total	12196

2) 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 object's 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.

C. Evaluation

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

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

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

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

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

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

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

III. Submissions and Results

There are eight teams participating in our campaign. We received full reports from five teams [3], [4], [5], [6], [7] and two short descriptions from two teams. The other one did not send us any papers. 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.

A. Methods and Features

The methods used by participating systems are presented in Table II. 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 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 [4].

TABLE II
Methods of 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-VietSentiWordNet
sa3	Ensemble: - 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, n-gram (1, 2 ,3) extracted syllables, n-gram (1, 2 ,3) extracted 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

In term of features, almost all systems use the basic n-gram features. TF-IDF also plays an important role in many systems [4], [5], [6]. In addition, some systems use external dictionaries of sentiment words, booster words, reverser words and emotion words to enrich their feature sets and help to gain better results [3], [7].

B. Results

The best results of all teams are reported in Table III 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 II. 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 term of F1 score. This strongly surprised us. The result of **sa1** is also much better than others’. We 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.

TABLE III
RESULTS OF THE PARTICIPATING SYSTEMS (%)

Team	Positive			Negative			Average
	P	R	F1	P	R	F1	F1
sa1	75.85	89.71	82.20	79.88	76.00	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	72.00	70.00	60.59	70.29	65.08	67.54
sa5	69.06	71.43	70.23	65.67	62.86	64.23	67.23
sa6	71.80	70.57	71.18	67.10	59.43	63.03	67.11
sa7	71.00	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

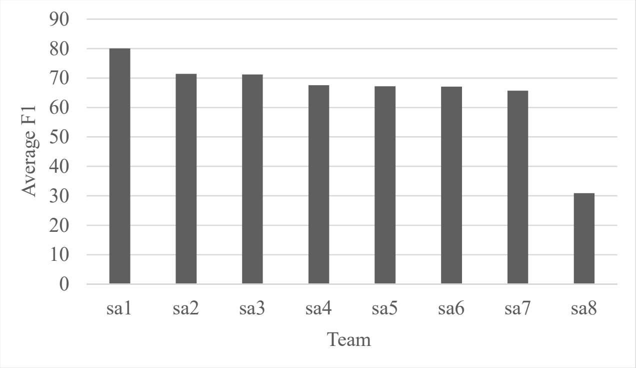

Fig. 1. Average F1 comparison.

The team **sa2** [6] 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 MLNN’s. The ensemble system of **sa3** [4] 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** [3] used SVM as learning method combining with n-gram features and various other features extracted from external dictionaries that help to gain average F1 score at 67.54%. Next, the report of team **sa5** [5] 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.

IV. CONCLUSION

We have described a new task in VLSP 2016 on Sentiment Analysis on Vietnamese texts for the first time. This first task attracted a good number of participants: 8 teams. All participating systems implement popular machine learning approach and also have many rich features to resolve the task.

In the campaign, the first dataset used to benchmark sentiment analysis systems for Vietnamese has been released. In fact, the dataset is quite simple because it only covers comments about just one object. However, we strongly believe that it will help to impulse the development of researching on this topic in the near future. In the next campaign, we hope that the new dataset will contain more complex cases, such as a review or comment can contain multiple objects and aspects with different sentiments. We also need other task such as aspect-based sentiment analysis.

ACKNOWLEDGMENT

We would like to thank the sponsors for helping financially to the construction of the datasets for the sentiment analysis task. And special thanks to all the teams who have participated and contributed to the success of this evaluation campaign.

REFERENCES

[1] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,” *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[2] 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 (forthcoming)*, 2016.

[3] 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.

[4] P. Minh Quang Nhat 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.

[5] T. Thy Thy, H. Xanh, and N. Nhung T.H., “A multi-layer neural network-based 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.

[6] 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.

[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.