An Ensemble of Shallow and Deep Learning Algorithms for Vietnamese Sentiment Analysis

Hoang-Quan Nguyen
Faculty of Information and Technology
Le Quy Don University
Ha Noi, Viet Nam
nguyenhoangquan@tcu.edu.vn

Quang-Uy Nguyen*
Faculty of Information and Technology
Le Quy Don University
Ha Noi, Vietnam
quanguyhn@gmail.com

***Abstract*—Sentiment analysis also known as opinion mining refers to the use of natural language processing to systematically identify and categorize opinions expressed in a piece of text. Recently, deep learning with ensemble techniques has achieved state of the art results in sentiment analysis. However, this approach has not been studied for Vietnamese corpus. In this paper, we propose an ensemble method by combining traditional (shallow) and deep learning algorithms. We tested our method on three Vietnamese sentiment datasets. The Experimental results showed that these approaches improve the accuracy of sentiment classification when compared to both individual deep and shallow algorithms.**

***Index Terms*—Sentiment analysis, deep learning, ensemble learning.**

I. Introduction

Sentiment analysis is the task of determining users opinion about products, movies, events or policies etc. in a piece of text. Predicting user's sentiment is of great important in many real world applications. The public interest is the main factor that effects the profit of products like movies, books, etc. Subsequently, this problem has received a great attention from researchers in recent years. For a comprehensive survey of sentiment analysis and opinion mining, reader are referred to [1]. The major tasks in sentiment analysis include:

- Subjective classification: aims to classify subjectivity and objectivity documents.
- Polarity sentiment classification: aims to classify an subjectivity document into one of two classes ("positive", "negative") or three classes ("positive", "negative" and "neutral").
- Rating: aims to rate the documents having personal opinions from 1 star to 5 stars (very negative to very positive).

Sentiment classification is the most important task among the above sentiment analysis tasks. In this paper, we focus on the polarity sentiment classification. Specifically, we attempt to classify the document into three classes ("positive", "negative" and "neutral"). Generally,there are two common techniques for sentiment classification: machine learning based approaches and lexicon based approaches [2]:

- Lexicon based approach: This technique includes dictionary and corpus based approach. Dictionary based approach often uses an existing dictionary which contains a collection of opinion words along with their positive or negative sentiment strength. The opinion words in the dictionary are then used to categorize the sentiment of the testing document. Corpus based approach often relies on using very huge amount of documents like Google search, Alta Vista search etc. The probability of occurrence of a sentiment word in conjunction with positive or negative set of words is estimated by performing a search on these corpus. Finally, the sentiment of the document is decided based on the estimated probability of the sentiment words.
- Machine learning based approach: Machine learning have become the main approach in sentiment analysis. Machine learning can be further divided into supervised and unsupervised approaches. For supervised approaches, we need two sets of annotated data for training and testing learning model. Recently, deep learning is the state of the art ma-

*Corresponding author

978-1-5386-7983-8/18/$31.00 ©2018 IEEE

chine learning techniques that offers the method for learning features representation in a supervised or unsupervised fashion [3].

In recent years, ensemble techniques have been used with deep learning algorithms [4] to improve the accuracy of sentiment classification. Ensemble learning is machine learning paradigm where multiple learners are trained to solve the same problem. While ordinary machine learning approaches that try to learn one hypothesis, ensemble methods try to construct a set of hypothesis and combine them for the prediction step [5]. The combination of ensemble techniques with deep learning has been proven to enhance the effectiveness of sentiment classification for several popular English sentiment datasets [6].

In this paper, we propose an ensemble of shallow (traditional) and deep learning algorithms for sentiment classification in Vietnamese language. To the best of our knowledge, this approach has not been studied for Vietnamese sentiment analysis. The experimental results showed that our method improves the accuracy of sentiment classification compared to the individual versions. The remainder of this paper is as follows. Section II briefly presents previous work on both ensemble techniques and deep learning approaches for sentiment analysis. Section III describes two main deep learning models used in this paper whereas Section 4 details the proposed ensemble model. In Section V, we describe the designed experimental setup. Experimental results are presented and analyzed in Section VI. Finally, Section VII draws some conclusions from previous results.

II. Related work

In this section we offer a brief summary of the previous research on ensemble methods and deep learning algorithms for Sentiment Analysis.

A. Deep learning for sentiment analysis

Deep learning has emerged as a powerful machine learning technique that learns the representations or features of the data and produces state-of-the-art prediction results. In sentiment analysis, deep learning is also popularly used in recent years. Most study focuses on using deep learning models to learn the text features automatically [7]. Continuous representations of words as vectors has proven to be an effective technique in sentiment analysis. Yoon Kim et al [8] used word's vectors of Google pre-trained models to extract features of sentiment sentences. The features are used as the input for a Convolutional Neural Network (CNN) to classify the sentiment of the sentences. Kai Sheng Tai et al. [9] proposed several variants of Long Short Term Memory (LSTM) network to tackle sentiment classification. All these models used word2vec [10] features as input values.

In Vietnamese documents, Kieu and Pham [11] proposed a rule-based system for Vietnamese sentiment classification in computer product reviews. Duyen [12] used machine learning to examine the impact of the score of a review on classifying the sentiment of a sentence. The author showed that machine learning has several advantages over a rule-based approach in sentiment analysis. Phan and Cao [13] used Skip-gram word estimation model with SVM-based classification for opinion mining Vietnamese food places text reviews. Recently, Vo et al. [14] proposed a multi-channel deep learning model for Vietnamese sentiment analysis. They combined the features of a CNN and LSTM network and use this feature a the input fo a softmax function to make the final prediction.

B. Ensemble learning

In ensemble learning, the main idea is to combine a set of models (base classifiers) in order to obtain a more accurate and reliable model. There has been a number of studies that used ensemble techniques for sentiment analysis. Xia et al [15] used multiple features include part-of-speech(POS), word relation (n-gram) and term frequency - Inverse document frequency (TF-IDF) in machine learning models. They also used the ensemble of three popular machine learning approaches including Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) for sentiment analysis. In another study, Nadia Felix et al [4] combined SVM, NB and DT models. They also used the ensemble features of bag-of-word (BoW), feature hashing and lexicon feature on two-class dataset from Twitter.

Recently, deep learning models with the word vector feature has also used for sentiment classification. Oscar Araque et al [6] proposed ensemble learning model using Recurrent Neural Network (RNN) with surface features, generic automatic word vectors and affect word vectors specifically trained for the sentiment analysis task. Surface feature includes a set of traditional text features extracting using SentiWordnet [16]. Overall, deep learning and ensemble techniques have received a great attention in sentiment analysis. In this paper, we will propose an ensemble model of deep learning and shallow algorithms for Vietnamese sentiment analysis.

The detailed description of our method is presented in Section IV.

III. Background

This section briefly describe two deep learning networks (CNN and LSTM) that are used for determining the sentiment of documents in this paper.

A. Convolution Neural Network

We used method similar to [8] in which each document is represented as a matrix $d \times s$ where s is the document length and d is the dimension of each word vector. Let $x_i \in \mathbb{R}^k$ be the k-dimensional word vector corresponding to the i^{th} word in the document. The length of the text is n which is expressed as follows:

$$x_{1:n} = x_1 \oplus x_2 \oplus ... \oplus x_n \qquad (1)$$

where $\oplus$ is the concatenate operator. In general, let $x_{i:i+j}$ refer to the concatenation of words $x_i, x_{i+1}, ..., x_{i+j}$. A convolution operation involves a *filter* $w \in \mathbb{R}^{hk}$, which is applied to a window of h words to produce a new feature. For example, a feature c_i is generated from a window of words $x_{i:i+h-1}$ by:

$$c_i = f\left(W\dot{x}_{i+h-1} + b\right) \qquad (2)$$

Here $b \in \mathbb{R}$ is a bias term and f is a non-linear function such as the hyperbolic tangent. This filter is applied to each possible window of words in the document $\{x_{1:h}, x_{2:h+1}, ..., x_{n-h+1:n}\}$ to produce a *feature map*.

$$c = [c_1, c_2, ..., c_{n-h+1}] \qquad (3)$$

with $c \in \mathbb{R}^{n-h+1}$. We then apply a max pooling operation over the *feature map* and take maximum value $\hat{c} = max\,\{c\}$ as the feature corresponding to this particular filter. The model uses multiple filters (with varying window sizes) to obtain multiple features. These features are passed to a fully connected softmax layer whose output is the probability distribution over labels.

B. Long Short Term Memory

Long Short Term Memory (LSTM) is the special kind of recurrent neural network. LSTM cells were introduced by Hochreiter [17] and were created in order to be able to learn long time dependencies. A LSTM cell comprises at time t, a state C_t and output h_t. Inside the LSTM, the computations are defined by doors that allow or not allow the transmission of information. These computations are governed by the following equation:

$$\begin{aligned}
u_t &= \sigma(W^u h_{t-1} + I^u x_t + b^u) \qquad (4)\\
f_t &= \sigma(W^f h_{t-1} + I^f x_t + b^f)\\
\tilde{C}_t &= tanh(W^c h_{t-1} + I^c x_t + b^c)\\
C_t &= f_t \odot C_{t-1} + u_t \odot \tilde{C}_t\\
o_t &= \sigma(W^o h_{t-1} + I^o x_t + b^o)\\
h_t &= o_t \odot tanh(C_t)\\
y_t &= softmax(W h_t + b)
\end{aligned}$$

Notice that at the final layer, a softmax function is used to make the final prediction.

IV. Methods

The structure of our ensemble model is presented in Figure 1. The model includes three steps. The first step is feature extraction, the second step is learning and the last step is ensemble step.

A. Features Extraction

In the first step, the document is preprocessed using stop word, special characters removing and then words tokenizing. After that, two techniques for extracting features are applied. For deep learning models, we used word2vec and for shallow algorithms, we used a popular feature, TF-IDF.

- Term Frequency and Inverse Document Frequency (TF-IDF) is a value that determines the weight of words in specified document of given text set [18].

  $$TFIDF(t,d,D) = TF(t,d) * IDF(t,D) \qquad (5)$$

  Where TF of term t is determined:$TF(t,d) = \frac{f(t)}{\sum_{t' \in d} f(t')}$ and IDF: $IDF(t,d,D) = log\frac{N}{|\{d \in D : t \in d\}|}$ Denote: t, t':term, d:document, D:document set, $N = \|D\|$
- Word2vec: Since there has not been any published pre-trained model for Vietnamese word2vec, we have trained our word2vec model for the experiments in this paper. We collected the Vietnamese text from Wiki at [1]. This dataset includes 33.783 documents in 10 subject and occupies 441 MB disk volume. We used skip-gram model and each word has dimensionality of 300. After training, we achieved a word vectors model of 38402 words.

[1] https://github.com/magizbox/corpus.viwiki/tree/master/viwiki

B. Learning algorithms

In the second steps, machine learning algorithms are used for learning from the features extracted in the first step. More precisely, two deep learning algorithms (CNN and LSTM) are applied to the word vectors and two shallow algorithms (Support Vector Machine-SVM and Logistic Regression - LR). For each deep learning model (CNN and LSTM), we tested three following variants:

- CNN/LSTM-rand: the baseline models where all words are randomly initialized and then modified during training.
- CNN/LSTM-static: the models with pre-trained vectors from word2vec. All words including the unknown ones that are randomly initialized, are kept static and only the other parameters of the model are learned.
- CNN/LSTM-non-static: Same as above but the pre-trained vectors are fine-tuned for each task.

The output of each learning algorithm is a vector where each value present the probability of the input document belonging to the corresponding class.

C. Ensemble Techniques

Based on the probability vector at the output, the algorithms are combined together using three following methods [19]:

- Average rule: calculates the average target probability value for each element in the classifiers then chooses the maximum value for labeling that element. Let c is number of classes. m is number of classifiers Label of k^{th} sample is predicted:

$$l^k = \underset{j=1..c}{argmax}\left(\frac{1}{m}\sum_{i=1}^{m} P_{ij}^k\right) \tag{6}$$

- Max rule: compare the results of the probability values by the target class of each element and then get the maximum results for labeling:

$$l^k = \underset{j=1..c}{argmax}\left(\max\{P_{ij}^k\}\right) \tag{7}$$

- Voting rule: the results of classification for each method will determine the label for each sample, the final label value is assigned based on mostly number.

$$l^k = \underset{j=1..c}{argmax}\left(\sum_{i=1}^{m} I\left(\underset{j=1..c}{argmax}\left(P_{ij}^k\right) = j\right)\right) \tag{8}$$

V. Experimental Settings

This section presents the parameter's setting of the algorithms and the datasets for testing the proposed ensemble technique.

A. Parameters Settings

For two shallow algorithms, LR and SVM, we used their default settings in *Scikit-learn* library. The settings for two deep learning models (CNN and LSTM network) are as follows:

- Long Short Term Memory: One hidden layer of 128 nodes is used. Each document being adjusted to a length of 50 words. If a document length is greater than 50, it is truncated. If the length is smaller than 50, it is padded with zero. Every word is a 300-dimension pre-trained vector.
- Convolutional Neural Network: In the convolution layer, we used one-dimension convolution methods from input values with 50 filter, filter size is [2x3]. In the next layer, a max pooling is used to reduce the data dimension. Next, the data are flatted to vectors and then concatenated. In the fully connected layer, we also used one hidden layer with 128 nodes, activation function is ReLU (Rectifier Linear Unit).

For both deep learning algorithms, the training was performed in 30 epochs and the best model on the validation set was recorded for making the prediction on the testing data.

B. Datasets

We tested the ensemble approach on three Vietnamese sentiment datasets.

- Vietnamese sentiment dataset (DS1), Table I. This three-class dataset is provided by Vietnamese Language and Speech Processing [2].

Table I: Vietnamese sentiment dataset from VLSP

	Positive	Neutral	Negative	Summary
Train	1700	1700	1700	5100
Test	350	350	350	1050
Total	2050	2050	2050	6150

- Vietnamese Sentiment food reviews (DS2), Table II. This binary classes dataset was collected from comments on *https://www.foody.vn* website and published at *https://streetcodevn.com/blog/dataset*.

[2] http://vlsp.org.vn/resources-vlsp2018

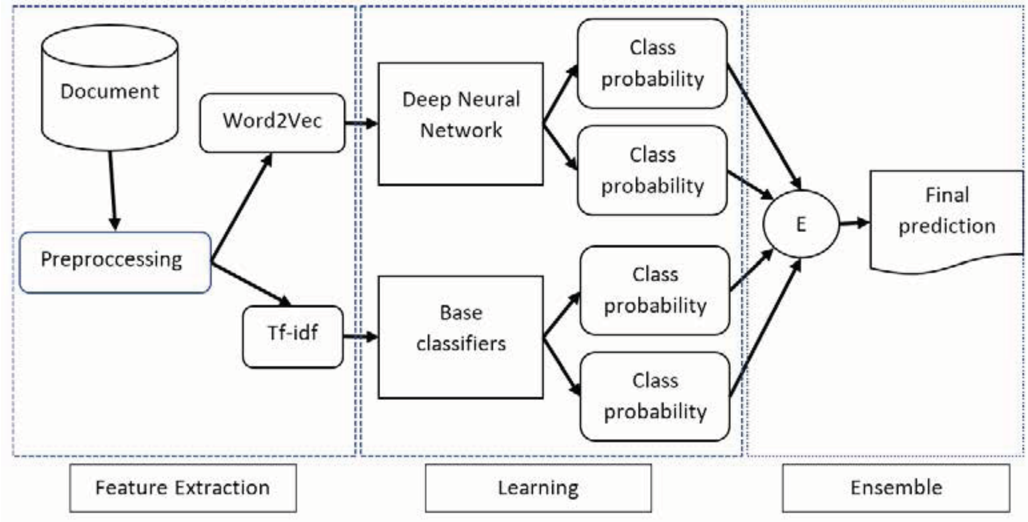

Figure 1: Ensemble model

Table II: Vietnamese sentiment dataset from foody.vn

	Positive	Negative	Summary
Train	15000	15000	30000
Validation	5000	5000	10000
Test	5000	5000	10000
Total	25000	25000	50000

- Vietnamese sentiment dataset reviews about restaurants and travel services from *http://tripnow.vn* website (DS3), Table III. This is binary class dataset: positive label ($if score > 5$) and negative label ($if score \leq 5$).

Table III: Vietnamese Sentiment dataset from tripnow.vn

	Positive	Negative	Summary
Train	4525	282	4807
Test	925	75	1000
Total	5450	357	5807

It should be noted that, since DS1 and DS3 datasets do not have validation set, so we have split the train set into training and validation set with ratio 1:10.

VI. Results and Discussion

For comparing classification algorithm, $accuracy$ is the most common criteria [20]. Formally, the accuracy of a method on a data set of n samples is calculated as follows:

$$MR = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i = Z_i\right) \tag{9}$$

where Y_i real label, Z_i predicted label, $I\left(\cdot\right)$ indicator function

We divided our experiments into two sets. In the first set, we combine the probability of only six of deep learning models and compare the result with the individual versions. In the second set, we compared the ensemble of shallow and deep models with all tested individual models.

The results of the first experiment are presented in Table IV [3]. It can be seen that, the ensemble techniques often help to improve the accuracy of the sentiment classification. On two first datasets (DS1 and DS2), three ensemble approaches are often better than the individual versions from 2% to 4%. On the last dataset (DS3), the ensemble approaches are only slightly better than the individual ones. The reason could be that, DS3 is an imbalanced dataset and the probability value of all individual algorithms for the major class is often much greater (close to one) than the value for the minor class. Subsequently, the ensemble of algorithms did not change the prediction label of the individual algorithms.

The results of the second experiment are presented in Table V. This table shows that combining both shallow

[3] The best value in Table V and Table IV is printed bold face.

Table IV: Comparison of ensemble techniques and with deep learning models.

Method	Feature	DS1	DS2	DS3
LSTM-non-static	word2vec	67.05%	85.28%	92.50%
LSTM-static	word2vec	67.05%	85.27%	92.50%
LSTM-rand	word2vec	60.76%	84.59%	92.50%
CNN-non-static	word2vec	63.81%	87.95%	92.20%
CNN-static	word2vec	62.10%	85.82%	92.60%
CNN-rand	word2vec	66.10%	88.11%	92.50%
Mean-Rule		**69.43**%	**88.78**%	92.50%
Max-Rule		67.62%	88.45%	92.50%
Vote-Rule		68.10%	88.14%	**92.70**%

and deep learning algorithms improves the accuracy to a greater extent. It can be seen that, the value of three ensemble approaches in Table V is alway higher than that value in Table IV about 1%. Comparing between three ensemble techniques, two tables show that Mean-Rule and Vote-Rule are often slightly better than Max-Rule.

Table V: Comparison of ensemble techniques and with all individual models.

Method	Feature	DS1	DS2	DS3
SVM	TF-IDF	68.48%	88.72%	93%
LR	TF-IDF	68.57%	88.37%	92.50%
LSTM-non-static	word2vec	67.05%	85.28%	92.50%
LSTM-static	word2vec	67.05%	85.27%	92.50%
LSTM-rand	word2vec	60.76%	84.59%	92.50%
CNN-non-static	word2vec	63.81%	87.95%	92.20%
CNN-static	word2vec	62.10%	85.82%	92.60%
CNN-rand	word2vec	66.10%	88.11%	92.50%
Mean-Rule		**69.71**%	89.13%	92.60%
Max-Rule		68.00%	88.62%	92.50%
Vote-Rule		69.33%	**89.19**%	**92.80**%

VII. Conclusion

In this paper, we proposed an ensemble model of shallow and deep learning algorithms and applying this model for Vietnamese sentiment analysis. We tested the ensemble methods on three popular sentiment datasets. The experimental results show that the ensemble models improve the accuracy of the individual algorithms in Vietnamese sentiment classification. Particularly, combining both shallow and deep learning algorithms is often better than the ensemble of only deep learning models.

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] J. Singh, G. Singh, and R. Singh, "A review of sentiment analysis techniques for opinionated web text," *CSI transactions on ICT*, vol. 4, no. 2-4, pp. 241–247, 2016.

[3] L. M. Rojas-Barahona, "Deep learning for sentiment analysis," *Language and Linguistics Compass*, vol. 10, no. 12, pp. 701–719, 2016.

[4] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka Jr, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170–179, 2014.

[5] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision support systems*, vol. 57, pp. 77–93, 2014.

[6] O. Araque, I. Corcuera-Platas, J. F. Sanchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.

[7] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1253, 2018.

[8] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[9] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

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

[11] B. T. Kieu and S. B. Pham, "Sentiment analysis for vietnamese," in *Knowledge and Systems Engineering (KSE), 2010 Second International Conference on*. IEEE, 2010, pp. 152–157.

[12] N. T. Duyen, N. X. Bach, and T. M. Phuong, "An empirical study on sentiment analysis for vietnamese," in *Advanced Technologies for Communications (ATC), 2014 International Conference on*. IEEE, 2014, pp. 309–314.

[13] D.-H. Phan and T.-D. Cao, "Applying skip-gram word estimation and svm-based classification for opinion mining vietnamese food places text reviews," in *Proceedings of the Fifth Symposium on Information and Communication Technology*. ACM, 2014, pp. 232–239.

[14] Q.-H. Vo, H.-T. Nguyen, B. Le, and M.-L. Nguyen, "Multi-channel lstm-cnn model for vietnamese sentiment analysis," in *Knowledge and Systems Engineering (KSE), 2017 9th International Conference on*. IEEE, 2017, pp. 24–29.

[15] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138–1152, 2011.

[16] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[19] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.

[20] M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University, Corvallis*, vol. 18, 2010.