Sentiment Analysis for Vietnamese using Support Vector Machines with application to Facebook comments

Vi Ngo Van
Data Mining Team
Big Data Department, Admicro
VCCorp
Hanoi, Vietnam
Email: vingovan@admicro.vn

Minh Hoang Van and **Tam Nguyen Thanh**
School of Information and
Communication Technology
Hanoi University of Technology
Hanoi, Vietnam
Email: {hoangminh.it.hut,mrtamb9}@gmail.com

***Abstract*—Sentiment Analysis covers the area of research that studies people's opinions, sentiments, evaluations, attitudes and emotions from written text. This has become one of the most active research fields in Natural Language Processing, Text Mining and Data Mining in general. In recent years, with the rapid growth of World Wide Web, there has been an exponential increase in the number of texts available online. These days, people tend to express more and more opinions about various kinds of subjects in their lives. This communication of sentiments may have a great influence on others' decisions via social networking services, such as Facebook, Twitter and so forth.**
In this paper, we present an effective Sentiment Analysis model for Vietnamese that is based on Support Vector Machine, an advanced Supervised Learning technique.
Keywords: Natural Language Processing, Text Mining, Sentiment Analysis, Vietnamese, Support Vector Machine, n-gram, Facebook, Supervised Learning.

I. Introduction

Facebook is one among the largest, most popular social networks in Vietnam. It has attracted a huge number of netizens who update their status, share their feelings, post comment and like others' postings on a daily basis. On this network, people also tend to give their either positive or negative reviews on products and services they have experienced. Therefore, the employment of information shared on Facebook surely helps companies to gather feedbacks from their customers. It is also useful when a company wants to do market survey or to learn more about the products of the rivals in order to adjust its own business strategy.

In a conventional way, customers' reviews are collected and analyzed completely by hands. It means that there should be a staff that are responsible for reading and grouping the feedbacks from different info channels. This job is obviously extremely tedious and requires a tremendous amount of time. When it comes to Facebook reviews, the task becomes impractical. Therefore, automating collecting and analyzing and classifying large amount of sentiment information become a must for companies professionally involved in data competitions.

However, fulfilling this goal is not easy. It is to say that doing sentiment analysis for formal text is difficult, doing sentiment analysis for informal text is even more challenging. We could name some of these challenges. First of all, Facebook postings and comments are often of variable lengths, some are short, some are long. This diversity makes it hard to apply techniques that require strict standardization. Second of all, informal language does not follow standard grammatical rule, i.e. no capitalized beginning letter, no punctuation, using abbreviations. Finally, sarcasms and ironies seem to be the hardest part to overcome.

Building a complete Customer Feedback System requires many processing steps from gathering information to exporting the final statistical report. However, in the scope of this paper, we focus only on the main part, opinion sentiment analysis. The core technique employed in our model is Support Vector Machines (SVM). This advanced classification method has demonstrated its strength in various Machine learning application.

II. Related work

The existing methods applied to sentiment analysis task can be grouped into several classes as machine learning, lexicon-based, statistical and rule-based approaches. The rule-based approaches seek for words in a text that bring opinion or sentiment meaning and then classify based on the number of these positive and negative words. The lexicon-based approaches tend to compute sentiment polarity for a text according to the semantic orientation, a measure of subjectivity and opinion, of words. Statistical models tries to find the head terms, link them to the true aspects and sentiment them into ratings. Machine learning techniques employ sentiment knowledge in training dataset to give predictions. With the fast growth in the field of ML study, many strong tools have been applied to solve the problems of sentiment analysis. Since sentiment analysis is merely a text classification problem, any existing supervised learning method can be applied, e.g., Naive Bayes classifi-

cation or Support Vector Machines (SVM) (Joachims, 1999; Shawe-Taylor and Cristianini, 2000). The first paper to take these approaches to classify movie reviews into two classes, positive and negative was Pang et al. (2002). The sentiment analysis model presented in this paper is also motivated by this approach.

III. Sentiment analysis model

A. Model preparation

1) Lexicon-sentiment list: As noted above, sentiment classification is barely a text classification problem by its nature. Conventional text classification primarily groups documents of different topics, e.g., politics, society, sciences, education and sports. To achieve that goal, the key features are topic-related words. In contrast, in sentiment classification problem, sentiment and opinion words that indicate positive or negative are more important. e.g., *tốt, xấu, đẹp, tuyệt, tồi* etc. In these examples, *tốt, đẹp, tuyệt* are positive sentiment words and *xấu, tồi* are negative sentiment words. Most sentiment words are adjectives and adverbs, but nouns (e.g., *rác*) and verbs (e.g., *yêu, ghét*) can also be used to express sentiments. Apart from individual words, there are also sentiment phrases and idioms. Again, apart from sentiment words and phrases, there are also other expressions or language compositions that can be used to express or imply sentiment and opinions. This should be listed in the full lexicon-sentiment list.

To build this word list, we exploit data from over 3 millions comments crawled from various Vietnamese forums and webpages. This set of data is trained by using word2vec, a model of neural network to learn word embedding, and finally passed through a manual review process.

2) Polarity reversers list: Negations, or more generally, polarity reversers, create inconsistent words which are a major cause of errors for polarity classification. Let us consider the example "Sản phẩm này không tốt". This sentence contains *"tốt"*, which is a positive-meaning word in lexicon-sentiment list. However, because of the preceding word *"không"* the sentiment turns into negative. *"không"* and similar words are commonly called as negation. Thus, the capture of these negation words is important while analyzing sentence sentiment.

3) Booster words list: Booster words, or degree modifier, are the words that impact sentiment intensity by either increasing or decreasing the intensity. Consider the examples "quá", "rất", "hơi", these words express different levels of sentiment in sentence and usually make the sentence become less neutral.

4) Emotion words list: Facebook postings are usually short, informal and contain many Internet slangs and emoticons. While this piece of information is helpful and usually used to improve sentiment analysis accuracy, it is sometimes hard to analyze the texts from complex topics that present sarcasms and ironies like social or political discussions.

5) Stop words list: Stop words are common words in a given language that do not bring any important meaning. They appear in almost every sentence of different contexts; thus, the removal of these words from processed data usually helps improving performance of sentiment models. One important thing that we should take care of is that having the list of given stop words, it requires to filter text again to keep meaningful stop words. For example, the stop words like *"không", "rất", "quá"* could be used for extracting important sentiment meaning.

B. Data preprocessing

The preprocessing stage of proposed SA model works as follows

- Word segmentation
- Removal of meaningless words
  - Proper nouns and proper noun phrases
  - Abbreviations
  - Stop words
  - Words that make non-sense in Vietnamese
- Extraction of uni-grams and bi-grams from text
- Determination of number and score of lexicon-sentiment words found in sentence. The rule is as follows:
  - If a sentiment word (positive or negative) is preceded by a booster word, we add score to sentiment polarity.
  - If a sentiment word (positive or negative) is preceded by a reversers, we add score to reversed sentiment polarity.
- Count emotion words.

Details about several steps of this preprocessing workflow are described below.

1) Segmentation: Before feeding raw data into the Sentiment Analysis model, we need to divide them into a sequence of component words. In English and several other languages using some forms of the Latin alphabet, the space is an ideal word delimiter. In Vietnamese, it is not words but syllables that are delimited, so a word could be composed by one or several tokens. There exists the case where a stand-alone token does not bring any meaning. According to this reason, a statistical learning tool for segmentation has been applied.

2) POS tagging: In the next step, part-of-speech (POS) tagging is the process of assigning a word to its right grammatical label, in order to truly perceive its role within the sentence. In the scope of Sentiment Analysis, we do not pay attention to the proper nouns. The words belonging to this class does not play any role in sentiment classification. Moreover, if we do not remove these words, they will naturally add more noise to the training set fed into SVM model.

Let us take an example, given the text about the infamous fly-in-the-bottle scandal of Tan Hiep Phat Beverage Group. If we do not remove the proper noun “Tan Hiep Phat” or proper noun phrase “ Tan Hiep Phat Beverage Group”, the model tends to learn and implicitly assign this key words to “negative” label which may bias the new coming results.

3) Stop words removal: Besides the mentioned above stop words and proper nouns, we also need to remove other meaningless tokens appeared in Vietnamese texts. These meaningless tokens are often composed of number or special characters. A simple way to accomplish this task is done by

using a full set of accepted signed and unsigned Vietnamese characters.

In the next part, we describe the core method using in our SA model.

C. Method

1) Support Vector machines: Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given training samples with classes, SVM tends to output the optimal hyperplane, a high-dimensional space, which can be used for categorizing new coming samples. Intuitively, optimal separation could be achieved by generating the hyperplane that has the largest distance to the nearest data point of any class. In almost case where the dataset is not linearly separable, SVM use kernel function to map a space of data points onto a new space in which data is linearly classifiable. A tutorial on SVM and its formulation could be found in Burges (1998) and Cristiani and Shawe-Tailor (2000). For details of the application of this model we refer to Joachims (2001).

2) n-gram and additional features: Machine Learning techniques do not work directly with raw text data until we could find a way to convert its textual content into numerical representation. In a simple manner, this representation could be done by encoding value of each word feature as its presence by 0 or 1, or TF-IDF score of that word. Depending on the number n of taken words, we may come up with coressponding so-called n-gram features.

An issue when using n-gram language models are out-of-vocabulary (OOV) words. This issue is encountered when the input includes words which were not present in our built model's dictionary. In such a scenario, the n-grams in the corpus that contain an OOV word are ignored and n-gram probabilities are still smoothed over all the words in the vocabulary even if they were not observed.

In addition to n-gram features, we could optionally utilize prior knowledge of Vietnamese lexicon-sentiment words. This is done by concatenating number of positive/negative words, sentiment score on positive/negative polarity and number of positive/negative emotion-words with original n-gram representation of input sentence. In experimentation section, we review the result using this additional information as input to SVM model.

IV. Experimentation

A. Experimental setup

When using Support Vector Machines or many other Machine Learning techniques for Sentiment analysis task, we need to consider defining an appropriate set of hyperparameters. The two important factors that may affect the classification performance of SVM are C and Gamma, the parameters for nonlinear SVM with Gaussian Radial Basis function kernel.

As solution to linear separable problem, standard SVM seeks to find a margin that separates all data points. However, for a nonlinear problem, this can lead to poorly fit models. This issue leads to the emergence of the concept "soft margin"

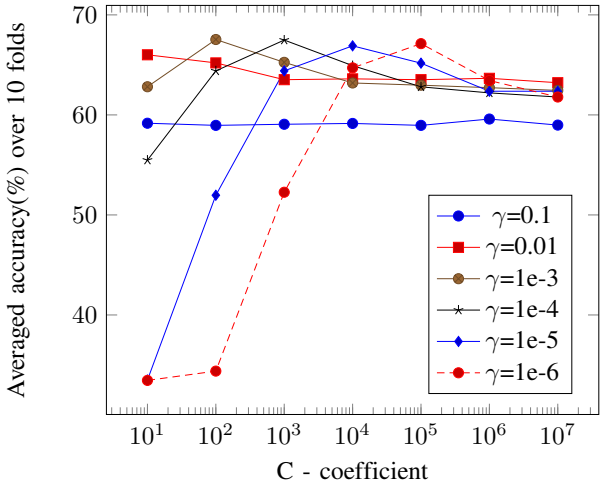

Fig. 1 Accuracy for validation

SVM that allows some data samples to be "ignored" or placed on the wrong side of the margin. Soft margin SVM helps to come up with a better overall fit while the impact of each individual support vector could be regularized by C, the parameter for the soft margin cost function.

Gamma plays a role as parameter handling non-linear classification. Let us say, when data is not linearly separable, we need to transform them to a higher dimension space or to raise them using nonlinear kernel function. For such a kernel function as Gaussian Radial basis function, Gamma is the free parameter to be adjusted.

There is a trade-off between setting large or small values of C and Gamma. In order to find the optimal hyperparameters setting of SVM model, we employ cross-validation framework, i.e. we first split the training data into folds and then do developing-and-validating model. The best values of C and Gamma are sought by looking at experimental result.

Training and test data contains 3 labels: positive, negative and neutral class, which brings neither positive nor negative meaning. When splitting data we make sure that percentage of each label class remains the same. This is to avoid causing bias in learning.

B. Results

1) Training and validation: This section shows the performance of our proposed sentiment analysis model on varying values of SVM hyperparameters. Having 10 folds splitted from given training data, we calculate classification accuracy for each fold and average these values. We set C values in range of 100, 1000, 1e4, 1e5, 1e6, 1e7 and Gamma values in range of 0.1, 0.01, 1e-3, 1e-4, 1e-5, 1e-6. The results are then depicted in Figure 1. As shown in this plot, the highest accuracy is achieved at C=100 and Gamma=0.001. This setting is fixed and used for the next part of experiment.

2) Experiment on test data: Having well-tuned hyperparameters from previous experiment, we use this optimal setting for learning Sentiment analysis model applied to the whole given training data. To show the effectiveness of the proposed method, we compare its result to other models. The first

TABLE I
F1 SCORES(%) COMPARISON

Labels -	SVM without lexicon-sentiment dictionary	SVM with lexicon-sentiment dictionary	Naive Bayes -
Positive	62.8	63.07	61
Negative	58.39	60.13	55.36
Neutral	43.01	41.86	44.95
Average	54.73	55.02	53.77

model used for that comparison is SVM without using lexicon-sentiment features. The second model is Naive Bayes, a simple but effective tool with small number of training samples. With these methods, we compute precision and recall values for each class label. These measures are then used to calculate average F1 scores for comparison.The final results are depicted in table I.

As shown in table, the proposed model of combining SVM with features derived from n-gram and additional lexicon-sentiment representation gains the best score, 55.03% in average. Model with SVM and pure n-gram features stays second with 54.73% and Naive Bayes scores 53.77%.

V. CONCLUSION

In this paper, an effective and complete model for Facebook Sentiment Analysis is presented. The model employs the strength of SVM in text classification along with building the features of words served for sentiment purpose.

In experimental part, the training data was firstly divided into 10 folds, which are used in cross-validation manner for developing and validating the model. We vary different settings of SVM to find the optimal hyperparameters. After that, we also conduct experiments to compare the results of SVM model with the baseline method using Naive Bayes. The experiments show our proposed model with n-gram and additional sentiment features performs the best in comparison to pure SVM model and baseline Naive Bayes. There is also to note that the average scores of all three models are not high. This could be caused of confusing data in neutral class, for which as a result we only gain the scores 43.01%, 41.86% and 44.95% respectively.

In general, sentiment analysis and opinion mining is still a challenging task. Sentiment analysis for Facebook and similar social interactive data is even more difficult. Unstructureness, large variance of length, are some among factors that may weaken many sentiment tools.

With recent achievements in Artificial Neural Network research, especially in the area of Deep Learning, there creates a big room for improving sentiment performance. Deep learning allows learning model to embed sentence structure and semantics well. These algorithm atempts to build representation of the entire sentence based on how the words are arranged and interact with each other, for example, *word2vec* and *paragraph vectors* have been shown to work very well. These methods are simple to train and implement. *Recurrent Neural Networks* like *LSTMs*, have proven to be able to gain very good performance. One important thing that should be in consideration while atempting to employ Deep Learning is to choose appropriate representation of input data for each applied method. Finally, in the top of this models, applying machine learning paradigms like ensemble learning may also help to improve the quality of our SA model.

In this work, to get sufficient data for building the dictionary of sentiment words, we use a set of manually-annotated text. For future work, we plan to improve the quality and coverage of this data source by automating the annotation process.

ACKNOWLEDGMENT

This research was supported by Admicro, VCCorp (Vietnam Communications Corporation). The authors would like to thank our colleagues from Data Mining team, who provided insight and expertise that greatly assisted the research. We also thank Mr. Tuan Hoang Anh, CTO at Admicro, for his comments and Mr. Bao Nguyen Chi, Mr. Thanh Luong for their tremendous assistance that greatly improved the manuscript.

REFERENCES

[1] J. Yi, T. Nasukawa, W. Niblack, R. Bunescu., *Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques*. In Proceedings of the 3rd IEEE international conferenc e on data mining (ICDM 2003), USA, pp. 427– 434,2003.

[2] T. Joachims: *Text categorization with support vector machines:learningwith many relevant features*. Proc. of ECML-98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE,pp. 137-142,1998.

[3] B. Liu, *Sentiment analysis and opinion mining. Synthesis lectures on human language technologies* , 5(1), pp.1-167.

[4] Y. Singh, P. K. Bhatia, and O. Sangwan, *A review of studies on machine learning techniques* International Journal of Computer Science and Security, vol. 1, no. 1, pp. 70–84, 2007.

[5] Mouthami, K., Devi, K.N. and Bhaskaran, V.M., *Sentiment analysis and classification based on textual reviews* In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE.

[6] Zhou, X., Tao, X., Yong, J. and Yang, Z., *Sentiment analysis on tweets for social events* In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on (pp. 557-562). IEEE, 2013, June.

[7] P.H. Shahana and B. Omman, *Evaluation of Features on Sentimental Analysis* Procedia Computer Science, 46, pp.1585-1592, 2015.

[8] Tripathy, A., Agrawal, A. and Rath, S.K., *Classification of Sentimental Reviews Using Machine Learning Techniques.* Procedia Computer Science,57,pp.821-829, 2015.