International Journal of Scientific and Technical Research in Engineering (IJSTRE)
www.ijstre.com Volume 8 Issue 6 ǁ Nov-Dec 2023.
ISSN: 2581-9941.

Use Filter Model To Feature Selection In Vietnamese Sentiment

Hoa Tran Thi Thieu
Ha Tinh University, Viiet Nam
tranthithieuhoa@gmail.com

Abstract: *Feature selection is an important issue for sentiment classification. To use machine learning methods in emotion classification, it is very important to choose the most suitable features for this task. In emotion classification, feature selection algorithms are a strategy to make text classifiers more efficient and accurate. Feature selection is a process of selecting a subset of features, where the optimality of the subset is measured by an evaluation criterion. Feature selection attempts to reduce the size considered in a task to improve performance on several dependent measures.*

For Vietnamese language, due to the features and specific characteristics of the Vietnamese language with diverse grammar, many polysemes, and different meanings in different contexts, these form objective and subjective causes in the classification process. In this study, the author use filter model have methods have been proved by experiment on the standard corpus of English, considering as the best to assess feature selection problem of sentiment classification in Vietnamese language.

Keywords *- Feature selection, Sentiment classification, sentiment analysis*

I. INTRODUCTION

Feature selection is an important issue for sentiment classification and it determines the features most relevant to the analysis process. The reason is that some words are more likely to correlate with class distribution than others. To use machine learning methods in sentiment classification, it is very important to choose the most suitable features for this task. In sentiment classification, feature selection algorithms are a strategy to make text classifiers more efficient and accurate. Feature selection is a process of selecting a subset of features, where the optimality of the subset is measured by an evaluation criterion. Feature selection attempts to reduce the size considered in a task to improve performance on several dependent measures. Vocabulary reduction is often used to improve the performance of the classification model. The high size of the feature space can be an impediment to the use of learning algorithms; it is important to reduce the feature space without affecting the performance of the analytic model. Some machine learning algorithms can work efficiently on data with low dimensionality, however, as the dimensionality of the data increases, they require more time or memory, making this process unfeasible. The native feature space consists of unique terms (token, n-gram, sentence, etc.) However, a moderately sized text collection can have tens to hundreds of thousands of objects. This is a very time-consuming computational task even with the most modern learning algorithms, resulting in high computational costs and long training times. The goal is to reduce the original space without sacrificing classifier performance, especially when the learner is automatically generated. Automatic feature selection methods may include removing non-informative statistical features or building new features based on statistical values of the entire text. Feature removal cannot be done simply by low frequency or high frequency feature removal. For example, technical features fall into specific categories, so their frequency is often low, even though they are effective keywords. Ideally, after applying feature selection, one should have a minimum sized feature subset, which is important for accurate analysis without sacrificing performance. Calculating the height is only one side of the problem. Over-equipping is another matter. Overfitting occurs when the analysis is correct in the domain of the training dataset, but it performs poorly outside that domain. Reducing feature space such as the removal of stop words and suffix classification reduces the risk of overfitting [8].

Feature selection also known as subset selection or variable selection is a commonly applied process in machine learning to solve the problem of height numbers. Its task is to select a subset of the important features and remove the extraneous, redundant, and noisy features for a simpler and more concise data representation. The benefits of feature selection are multiple. First, feature selection saves most of the learning curve's runtime by eliminating redundant and extraneous features. Second, the extraneous, redundant and noisy features do not interfere, the learning algorithm can focus on the most important data and build simpler but more accurate data models. Therefore, the performance of the classifier is improved. Third, feature selection can assist us in building a simpler and more general model and better understanding the basic concept of the task.

For Vietnamese, Vietnamese features include Vietnamese language parameters such as number of words, number of syllables, uni-gram, bi-gram information, .. specifically with the research has shown the Vietnamese parameters are as follows: number of words: 40.181 words. (The most used and widely used words); Number of syllables: 7.729 syllables. In which 81.55% of the syllables are also single words. 70.72% of compound words have 2 syllables. 13.59% of compound words have 3.4 syllables. 1.04% of compound words have 5 or more syllables. Thus, Vietnamese has a very large number of features, in order to classify sentiments, it is necessary to select appropriate features. In principle, it includes all words in the language, so with 40.181 words in Vietnamese, the number of spatial dimensions is very large, making the classification problem difficult to handle. In fact, the features must be selected in order to shorten the dimensionality of the feature space by removing the unimportant feature components but still ensure the accuracy of the text content [17].

II. MODELS FEATURE SELECTION

Feature selection methods can be mainly divided into three models: filter model, wrapper model and embedding model [16].

In the filter model, the features are selected before the induction step, thus independent of the learning method that will use the output of the feature selection. In addition, unrelated features are filtered (hence the name), so that the touch algorithm uses a subset of features that results in the best performance. These methods rank features according to some criteria and ignore all features that do not score enough. Due to its computational efficiency, filtering methods are very popular for data of high size. Some common filtering methods are F-score [6], mutual information [2], information acquisition [3] and correlation [13].

In the wrapper model, feature selection occurs outside of the underlying induction, but it uses the inductive algorithm as a subroutine instead of as a post-processor. Wrapping models involve a predefined learning model that selects features that measure the learning performance of a particular learning model [13]. While wrappers can produce better results, they are expensive to run and can break with a very large number of features. This is due to the use of learning algorithms in evaluating feature subsets every time.

Like the wrapper model, the embedded model depends on the classifier and allows for dependent features, but it has less computational complexity. In this model, feature selection occurs in the basic inductive algorithm, however, in this case the search for an optimal subset of features is integrated into the classifier structure and can be considered as a search in the associative space of feature and hypothesis subsets. The lack of 'tuning' between feature selection and classifier can lead to suboptimal results, however compromises must be achieved. A perfect selection of features will make the classification more efficient and accurate. The use of feature selection in text classification has been well studied and it has been shown that a sharp reduction in feature space does not show a significant reduction in performance, and sometimes even leads to an increase.

There are several methods of feature selection, and although one cannot assume that there is a perfect fit for any given task, some methods are more suitable than others. It depends on the size of the corpus and the applied machine learning method. Chen and Wu confirmed that a good feature selection method can make the simplest classifier model obtain satisfactory performance through training [4]. There are several studies that compare most metrics. However, the results are often inconsistent, possibly due to a number of issues. The most obvious is the data set selection. Data set characteristics can often weaken or strengthen the outcome of feature selection. Most feature selection methods involve the same basic process: assign weights to each feature, score features based on their weights, and retain only a specific number of features, or compute function is valid, within a pre-specified range. How feature selection and classifiers interact is key to achieving the best possible performance. This is especially true in the embedded and wrapped models, but even in the filter model, the selection of features from the feature space is very important. Although all three methods can be applied

separately, there are also some studies combining filter and wrapper [9]. Feature selection algorithms have been considered in [10]. For a large number of features, evaluating all states is computationally infeasible, requiring metaheuristic search methods. More recently, nature-inspired metadata algorithms have been used to select features, namely: seed swarm optimization [5], genetic algorithm-based attribute reduction [7], attractive search algorithm [14]. These methods try to achieve better solutions by applying knowledge from previous iterations.

III. METHODS FEATURE SELECTION

Here, we present some feature selection methods, these methods have been proved by experiment on the standard corpus of the world, considering English as the best. So the feasibility of this method is very high.

1. Chi-square Statistic

The Chi-square statistic measures the lack of independence between the text term t and the text category c and can be compared to the x^2 distribution with one degree of freedom to judge the extremeness. The mathematical definition of Chi-square is as follows [12]:

$$\chi^2(t_k, c_i) = \frac{|N|.[P(t_k, c_i).P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i).P(\bar{t}_k, c_i)]^2}{P(t_k).P(\bar{t}_k).P(c_i).P(\bar{c}_i)}$$

Where the quantities are defined as follows:

$P(t_k)$: probability of feature t_k

$P(c_i)$: probability of class c_i

$P(\bar{t}_k, c_i)$: probability upon random selection of document x; feature t_k will not appear in x, and x belongs to class c_i.

$P(t_k, c_i)$: conditional probability t_k belonging to class c_i

|N| : is the total number of documents.

2. Odds Ratio

The Odds Ratio measures the odds of a term t occurring in documents in a category c divided by odds of a term t not occurring in documents in category c. OR score for a term t and a category c is defined as follows [12]:

$$\mathrm{OR}(t_k, c_i) = \frac{P(t_k \mid c_i).(1 - P(t_k \mid \bar{c}_i))}{(1 - P(t_k \mid c_i).P(t_k \mid \bar{c}_i)}$$

There fore, choosing $J(\tilde{W}) = IG(t_k, c_i)$, $J(\tilde{W}) = MI(t_k, c_i)$, $J(\tilde{W}) = \chi^2(t_k, c_i)$, $J(\tilde{W}) = OR(t_k, c_i)$, $J(\tilde{W}) = GSS(t_k, c_i)$, we will get feature selection function in training space. Based on the criterion function $J(\tilde{W})$ above, we need to find the way to select the features that give the highest results of classification. In general, all functions listed above are connected locally to the class ci to assess the value of the term tk in all classes (ie it is independent of the class). In the experiment, we can use the function $f_{sum}(t_k) = \sum_{i=1}^{|C|} f(t_k, c_i)$ or the function with weighted sum $f_{wsum}(t_k) = \sum_{i=1}^{|C|} P(c_i) f(t_k, c_i)$ or maximum function $f_{\max}(t_k) = \max_{i=1}^{|c|} f(t_k, c_i)$ where $f(t_k, c_i)$ is the function $J(\tilde{W})$ above. All these functions aim to get the best features for the class ci although they are differently distributed in the training set.

3. Optimal Orthogonal Centroid Feature Selection

OCFS is the method having the same function as the 5 methods mentioned above but giving better results. OCFS method is based on the Orthogonal Centroid (OC) algorithm.

+ Orthogonal Centroid Algorithm

Orthogonal Centroid algorithm is a proposed supervised feature extraction algorithm which utilizes orthogonal transformation on centroid [15]. It has been proved very effective for classification problems on text data. OC algorithm is based on the Vector Space Computation in linear algebra by QR matrix decomposition. The OC algorithm also aims to find the transformation matrix $W \in R^{dxp}$ that maps each column $x_i \in R^d$ of $X \in R^{dxp}$ to a vector $y_i \in R^p$, Criterion $J(\tilde{W})$ is arg max J(W) = arg max trace (WTSbW), subject to WTW = I

where: $S_b = \sum_{j=1}^{c} \frac{n_j}{n}(m_j - m)(m_j - m)^T$

where: - m_j the mean vector of the j^{th} class is $m_j = (1/n_j)\sum_{x_i \in c_j} x_i$

- n_j is the class size of j:

- m the mean vector of all these documents is $m = (1/n)\sum_{i=1}^{n} x_i = (1/n)\sum_{j=1}^{c} n_j m_j$

+ OCFS Algorithm [15]

The solution space of the feature selection problem is discrete and consists of all matrices $\tilde{W} \in R^{dxp}$ that satisfy the constraint given above. the feature selection problem according to criterion $J(\tilde{W})$ is an optimization problem: $\arg\max J(\tilde{W}) = \arg\max trace(\tilde{W}^T S_b \tilde{W})$ subject to $\tilde{W} \in H^{dxp}$.

The elements in the formula are defined as in the OC section. Suppose $K = \{k_i, 1 \le k_i \le d, i = 1,2,...p\}$ is a group of indices of features. $\tilde{W}$ is a binary matrix with its elements of 0 or 1, and there are 1 and only 1 no 0 element in each column. On the other hand, we have:

$$trace(\tilde{W}^T S_b \tilde{W}) = \sum_{i=1}^{p} \tilde{w}_i^T S_b \tilde{w}_i = \sum_{i=1}^{p}\sum_{j=1}^{c} \frac{n_j}{n}(m_j^{k_i} - m^{k_i})^2$$

Thus, the nature of the OCFS issue is to look for the above set K as a maximum: $\sum_{i=1}^{p}\sum_{j=1}^{c} \frac{n_j}{n}(m_j^{k_i} - m^{k_i})^2$)

This motivates to propose an optimal feature selection algorithm according to $J(\tilde{W})$.

The details of the OCFS algorithm:

+ Input: Datasets (Copus)

+ Method:

- Step 1, compute the centroid mi, i=1, 2,…, c of each class for training data;

- Step 2, compute the centroid m of all training samples;

- Step 3, compute feature score $s(i) = \sum_{j=1}^{c} \frac{n_j}{n}(m_j^i - m^i)^2$ for all the features;

- Step 4, find the corresponding index set K consisted of the p largest ones in set S={s(i)| 1≤i ≤ d}

+ Out put: Recall Precision and Micro F1 of datasets.

+ Algorithm Analysis

- In terms of calculation: The OCFS algorithm complexity is O(cd). The OCFS algorithm is easy to be installed, faster in calculation and more effective than the 5 methods mentioned above.

- In terms of number of selected features: Without loss of generality, suppose the feature score of all the d features are s(k1) ≥ s(k2) ≥…≥ s(kd), the energy function is defined as:

$$E(p) = \frac{\sum_{j=1}^{p} s(k_j)}{\sum_{i=1}^{d} s(i)}$$

we can get the optimal number of features p by: p= argmin E(p) subject to E(p) ≥ T, where T >= 80%.

IV. VERIFICATION EXPERIMENT

To conduct the experiment, we edited the data that is the reviews of tourist destinations based on tourism websites, social networking sites [18-26]. Data collected from the Web has 1280 text files gathered for machine learning and testing. The data will be preprocessed, normalizing spelling, punctuation, removing spaces more than 1 space, input marks, blank lines, strange characters, etc. The whole is then edited and selected according to linguistic criteria for experimentation.

The experimental process was conducted in accordance with the sentiment classification process. First, the collected data was preprocessed text, then performed word segment. In our research experiments, we have used the Pointwise method [1] to solve the Vietnamese word segment problem. Next, we rely on the list of stopwords referred from the website to remove the stopwords.

The process selected features we use 3 different methods Chi-square (CHI), Odds Ratio (OR) and Optimal Orthogonal Centroid Feature Selection (OCFS) for experiment. Finally, we use the algorithm Support Vector Machines [11] to solve for the trained system. We use 70% of the acquired data as training data and use the remaining 30% as test data.

Results of sentiment classification with the SVM model by 3 methods of feature selection: OCFS, CHI and OR:

SVM model with 2500 selected features

Table 1: Results with 2500 features

Method of feature selection	Recall	Precision	F1
OCFS	93.01%	91.14%	92.07%
CHI	92.28%	88.31%	90.25%
OR	84.53%	80.53%	82.48%

SVM model with 5000 selected features

Table 2: Results with 5000 features

Method of feature selection	Recall	Precision	F1
OCFS	96.97%	92.24%	94.55%
CHI	95.70%	92.72%	94.19%
OR	94.41%	91.57%	92.97%

SVM model with 7500 selected features

Table 3: Results with 7500 features

Method of feature selection	Recall	Precision	F1
OCFS	95.94%	92.95%	94.42%
CHI	94.74%	92.71%	93.71%
OR	92.66%	82.20%	87.12%

V. CONCLUSION

The classification results obtained with SVM model by the three methods of feature selection of OCFS, CHI, and OR with 2500, 5000 and 7500 features show that with the case of 2500 features, the results are the lowest on all methods, results of the methods in the case of the features of 5000 is higher than that in the case of features of 7500. This indicates that with the features of 5000 the results are the best.

The classification results when using the OCFS feature selection method with the number of features 2500, 5000, 7500, respectively, are higher than when using the CHI method. The classification results when using the feature selection method CHI with features of 2500, 5000, 7500, respectively, are higher than the OR method. Thus, the classification results of the OCFS feature selection method with the number of features 2500, 5000, and 7500, respectively, give the best results among the three methods OCFS, CHI, OR.

The results of the OCFS feature selection method are higher than that of the CHI method. The results of the feature selection method CHI are higher than the OR method. Thus, the OCFS feature selection method gives the best results.

The classification results on the SVM models in the cases of features of 5000 and 7500 are higher than that with the features of 2500. The highest classification result is the one of SVM model with the selected features of 5000.

Besides, the features selected in the tests are appropriate (not too small and not too large). If the features are too small, they will not be good, and if they are too large, they will cause noise, making the classification results reduce or redundant, increasing the processing time of the program.

From the experimental results, we find that the feature selection method is effective. The evidence is confirmed through the results of selection by three methods. This once again confirms the feasibility of feature

selection methods applied in Vietnamese. That means that if the number of features in Vietnamese documents is selected appropriately, the classification results will be high, Choosing the right number of features will bring the best effect. This is completely consistent with the experimental results on the problem of document classification in English.

REFERENCES

[1] Luru Tuan Anh, Yamamoto Kazuhide. Applying the Pointwise method to the word segment problem for Vietnamese. Natural Language Processing Laboratory Department of Electrical Engineering Nagaoka University of Technology 940-2188, Nagaoka City, Niigata, Japan.

[2] L.T. Vinh, S. Lee, Y.-T. Park, B.J. d'Auriol, A novel feature selection method based on normalized mutual information, Appl. Intell. 37 (2010) 100–120.

[3] M.H. Aghdam, N. Ghasem-Aghaee, M.E. Basiri, Text feature selection using ant colony optimization, Expert Syst. Appl. 36 (2009) 6843–6853.

[4] Y. Chen and O. Wu. Semi-automated feature selection for web text filtering. In Fuzzy Systems and Knowledge Discovery (FSKD), 2010, *Seventh International Conference on, volume 6*, IEEE, 2010, pages 2513–2517.

[5] L.Y. Chuang, C.H. Yang, J.C. Li, Chaotic maps based on binary particle swarm optimization for feature selection, Appl. Soft Comput. 11 (2011) 239–248.

[6] S. Ding, Feature selection based F-score and ACO algorithm in support vector machine, in: Proceedings of the 2nd International Symposium on Knowledge Acquisition and Modeling, 2009.

[7] Alper Kursat Uysal, Serkan Gunal, Text classification using genetic algorithm oriented latent semantic features, *Expert Systems with Applications, Volume 41, Issue 13*, 1 October 2014, Pages 5938-5947.

[8] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization, *ACM Computing Surveys, Vol. 34*, No. 1, March 2002, pp. 1–47.

[9] Gunal, S. Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering & Computer Sciences*, 20, (2012). 1296–1311.

[10] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (4) (2005) 491–502.

[11] J. Platt, "Sequential minimal optimization: A fastalgorithm for training Support Vector Machines", Technical Report MSR-TR-98-14, MicrosoftResearch, 1998

[12] Forman, 2003): Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification, *Journal of Machine Learning Research, vol. 3*, pp. 1289—1305 (2003).

[13] M. Kabir, Md. Shahjahan, K. Murase, A new local search based hybrid genetic algorithm for feature selection, Neurocomputing

[14] E. Rashedi, H. Nezamabadi-pour, Feature subset selection using improved binary gravitational search algorithm, J. Intell. Fuzzy Syst. (2014), 26 (3) 1211–1221.

[15] Jun Yan-Ning Liu-Benyu Zhang-Shuicheng Yan. (2005) OCFS: Optimal Orthogonal Centroid Feature Selection for Text Categorization, Microsoft Research Asia, China.

[16] Uysal, A. K., & Gunal, S. A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, 2012. 226–235.

[17] Http://viet.jnlp.org/home

[18] https://phongnhakebang.vn

[19] Https://dulichtoday.vn

[20] https://mytour.vn

[21] http://dulichkhatvongviet.com

[22] https://www.tripadvisor.com.vn

[23] Https://www.tourismdanang.vn/diem-du-lich/ba-na-hills/

[24] Https://vinpearl.com/vi

[25] Https://www.vntrip.vn

[26] Https://www.klook.com