



Chinese-Vietnamese bilingual news unsupervised sentiment classification based on word-weighted JST algorithm

Siqi Lin, Zhengtao Yu, Shengxiang Gao, Junjun Guo and Yulong Wang

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

April 14, 2019

Chinese-Vietnamese bilingual news unsupervised sentiment classification based on word-weighted JST algorithm

Siqi Lin
School of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
E-mail: 852605791@qq.com

Zhengtao Yu *
School of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
*Corresponding Author
E-mail: ztyu@hotmail.com

Shengxiang Gao
School of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
E-mail: gaoshengxiang.yn@foxmail.com

Junjun Guo
School of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
E-mail: guojjgb@163.com

Yulong Wang
School of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
E-mail: 694575700@qq.com

Abstract—Sentiment classification is an important part of text sentiment analysis. Usually sentiment classification methods are supervised models or semi-supervised models, but well-labeled corpora are often difficult to obtain. In the Chinese-Vietnamese bilingualism, the lack of emotional resources leads to the accuracy of cross-language sentiment classification is not as accurate as that in a single-language environment. Unsupervised sentiment classification can improve the accuracy of sentiment classification by constructing emotional dictionary and incorporating word weights. In this paper, the Gibbs sampling process is firstly guided by the emotional polarity of the words in the sentiment dictionary, which helps the model to predict the vocabulary emotion and subject distribution. Then this paper proposes an unsupervised sentiment classification method based on the word weighted Joint Sentiment/topic (JST) model algorithm. Through the linear combination of three kinds of weighted methods, the emotional weight was integrated into the vocabulary. This approach improves the impact of emotional words in the Gibbs sampling process, and ultimately improves the accuracy of sentiment classification. Experiments show that the word-weighted JST model combined with prior knowledge has greatly improved the accuracy of Vietnamese sentiment classification.

Keywords—sentiment classification, unsupervised, emotional dictionary, weighting method, Chinese-Vietnamese

I. INTRODUCTION

Sentiment classification is one of the important contents of sentiment analysis and opinion mining. At present, there are two methods based on rules and statistics. The machine learning

method and the text representation model are two core contents of the statistical sentiment classification method. Machine learning methods include supervised, semi-supervised and unsupervised sentiment analysis. The training of classifiers in supervised and semi-supervised machine learning methods requires a certain number of labeled training samples. However, the manual tagging process is time consuming. Unsupervised machine learning does not require labeling for training. At present, there are two main unsupervised learning methods applied to sentiment classification: one is based on emotional dictionary. The method first scores the tendency of emotional words, and then determines the emotional polarity of the document according to the positive and negative scores. Another way is to incorporate emotional information into the text representation and then estimate the probability of emotional polarity in the document.

Cross-language text sentiment classification is using rich corpus in other languages to solve the problem of sentiment classification in another language. The difficulty in classifying cross-language texts is to establish the connection of words or texts between different languages. Zhou[1] proposed a cross-language sentiment classification model based on LSTM network. The model is characterized by LSTM document modeling in the source language and target language respectively. Then the paper introduces a hierarchical attention mechanism to capture the emotional features of words and sentences. Finally they complete the sentiment classification. Zhou[2] extracted the sentence-level emotional representation. Then they incorporate the sentence-level emotional representation into the semantic representation of the document. If we use RNN for bilingual document modeling, we need bilingual parallel documents. However, the Chinese and Vietnamese parallel documents are difficult to obtain. Therefore,

Fund Project: National key research and development plan project (Grant Nos.2018YFC0830105, 2018YFC0830100), National Natural Science Foundation of China (Grant Nos. 61732005, 61672271,61761026, and 61762056), Yunnan high-tech industry development project (Grant No. 201606), and Natural Science Foundation of Yunnan Province (Grant No. 2018FB104)

we consider using unsupervised methods to classify sentiment polarity. Chen[3] proposed a sentiment classification method based on topic seed words. The method obtains the topic seed words through $Frequency(tf)$ and $TF-IDF(tf_i)$. Then, they reconstructs the sentences with the extracted topic seed words to form a new topic text. Finally, they use the new reconstructed text to achieve the joint discovery of the theme and emotions. Ouyang[4][5] proposed a multi-granularity hybrid model based on LDA. The model uses the document distribution to generate local distribution, which increases the accuracy of local emotion/topic estimation, and thus improves the classification effect of the emotional theme model. Lin[6] proposed a joint sentiment theme model(JST) model for sentiment classification. Pan[7] proposed the sentiment classification of news based on JST algorithm. The method first extracts the emotional topic sentence of the news. Then they integrate emotional topic sentences into the JST model to classify the sentiment of the news. However, compared with supervised learning, the unsupervised model has a poor classification effect.

This paper first constructs an emotional dictionary in Chinese and Vietnamese. Then we propose an unsupervised sentiment classification model for Chinese and Vietnamese news. Due to the lack of emotional resources in Vietnamese, we use Chinese emotional resources to expand the Vietnamese emotional resources. Therefore, we build a bilingual dictionary of Chinese and Vietnamese. The word-weighted JST model in this paper is based on the traditional emotional theme model, the Joint Sentiment/topic Model. Since the JST model randomly assigns emotional labels to a certain word during the sampling iteration process, so we directly assign emotional labels to the words by emotional dictionary. In other words, we let the model guide the classification process by adding prior knowledge. In the JST model, each word of the corpus is equally important to the generated emotions and subject information. But in fact, words with emotional tendencies and words with domain characteristics have a strong influence on emotions and themes. Therefore, this paper integrates emotional weights into vocabulary by calculating the association between emotional seed words and other words in the corpus. This approach allows each word to be distinguished during the Gibbs sampling process and improves the accuracy of sentiment classification. Experiments show that the accuracy of sentiment classification of word-weighted JST model combined with prior knowledge has indeed improved.

II. CONSTRUCT BILINGUAL EMOTION DICTIONARY IN CHINESE AND VIETNAMESE

In the field of natural language processing, the sentiment dictionary is a very valuable basic resource. It is often used in the field of sentiment classification, emotion summaries[4-8]. However, because of the imbalance between Chinese emotional resources and Vietnamese emotional resources, the further research on Vietnamese sentiment analysis tasks is restricted. Therefore, we use Chinese emotional resources and news corpora to build Chinese and Vietnamese emotional dictionary.

Because of the language differences between Chinese and Vietnamese, we use graph models to represent bilingual emotional seed words. Then we use the label propagation algorithm and mutual enhancement algorithm to make emotional information spread between Chinese and Vietnamese.

The specific ideas are as follows: First, we extract the seed words from the unmarked Chinese and Vietnamese documents by the k-means algorithm. At the same time, we mark the emotional tendency of the seed word. Then we construct a Chinese and Vietnamese graph model for all the labeled emotional seed words. We use the word as the node of the graph model, and use the point mutual information(PMI) between the words to represent the distance between the graph nodes. We use the label propagation algorithm on the graph model to get the emotional estimate of the word. Finally, we use the mutual enhancement learning algorithm for the Chinese and Vietnamese graph models. The purpose of using the mutual enhancement algorithm is that we use the polarity estimation of Chinese words to improve the polarity prediction of Vietnamese words. At the same time, we use the emotional predictions of Vietnamese words in turn affects the emotional prediction of Chinese words. The idea of the mutual enhancement algorithm is to construct the connection between Chinese and Vietnamese through a bilingual dictionary. We first use the label propagation algorithm to continuously update the Chinese (Vietnamese) and Vietnamese (Chinese) emotional values obtained through bilingual dictionaries. Then, we continually iterate over the unmarked news documents and reconstruct the graph model of the bilingual sentiment words. Finally, we get a map of the emotional words in Chinese and Vietnamese, and get the domain emotional words in Chinese and Vietnamese according to the map.

We captured 4000 Chinese news from Sohu News, NetEase News and Tencent News as Chinese unlabeled data sets. We captured 8000 Vietnamese news as Vietnamese unlabeled data sets from the Vietnamese People's Daily, the Vietnam Communist Party, Vietnam Daily News, Vietnam Defense Network, *Ngôi Sao* and *Zing*. The scales of Chinese and Vietnamese sentiment lexicons constructed by the graph model are shown in Table I. There are 4626 emotional words in various fields of Chinese and 2939 emotional words in various fields of Vietnamese.

TABLE I. CHINESE AND VIETNAMESE EMOTIONAL DICTIONARY SCALE

Fields	Languages	
	Chinese	Vietnamese
military	1248	722
political	1377	771
Finance	1121	813
amusement	880	633

III. WORD WEIGHTED JST MODEL

A. JST Model

The Joint Sentiment/topic Model (JST) is a traditional emotional theme model. The JST model is an extension of the LDA model. It is an unsupervised classification model consisting of a four-layered structure of words, documents, topics, and emotional labels. We build three matrices through three links: the link between document and emotion, the link between topic and emotion, the link between word, topic and emotion. Then we use these three matrices to estimate the distribution of π , θ and ϕ . So we can get the proportion of the subject and the proportion of sentiment in each document. This model has been used extensively in sentiment classification

tasks for user reviews, such as movie reviews, book reviews, etc. Similarly, the joint emotional topic probability model can also be used to solve document-level sentiment classification tasks. Its graph model is shown in Fig. 1.

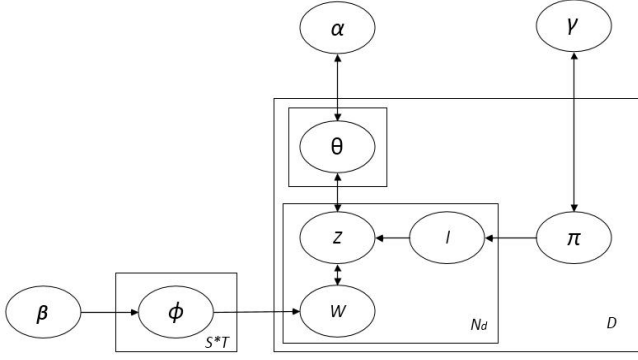


Fig. 1. Image representation of the JST model

The pseudo code for the Gibbs sampling procedure of JST[6] is shown in Fig. 2.

- 1: Initialization matrix ϕ (word-theme-emotion $V \times T \times S$), matrix θ (topic-emotion-document $T \times S \times D$), and matrix π (emotion-document $S \times D$)
- 2: **for** $m=1$ to M Gibbs sampling iteration do:
- 3: Read a word from the document, randomly assigned to the subject tag randomly.
- 4: Calculate the probability that the emotional label is k and the topic label is j based on
$$P(z_i = j, l_i = k | w, z_{-i}, l_{-i}, \alpha, \beta, \gamma) \propto \frac{\{N_{w_i, j, k}\}_{-i} + \beta}{\{N_{j, k}\}_{-i} + V\beta} \cdot \frac{\{N_{j, k, d}\}_{-i} + \alpha}{\{N_{k, d}\}_{-i} + \alpha} \cdot \frac{\{N_{k, d}\}_{-i} + \gamma}{\{N_d\}_{-i} + \gamma}$$
- 5: Reselecting a topic tag j based on the estimated probability in step 4;
- 6: Choose an emotional tag
- 7: Update matrix ϕ , θ and π with new sampling results
- 8: Go to step 2 until all words have been processed
- 9: **end for**

Fig. 2. The Gibbs sampling procedure of JST model

B. WORD WEIGHTED JST ALGORITHM

Due to the lack of emotional resources in Vietnamese, we expand the emotional resources of Vietnamese by using the Chinese and Vietnamese emotional dictionaries constructed by mutual reinforcement learning. Therefore, this paper uses the emotional dictionary to assign emotional labels to the vocabulary of each Vietnamese document. Finally, this paper conducts unsupervised sentiment classification based on the word weighted JST model.

In the process of sentiment classification of news documents, we should try to block all factors that affect the classification of emotions. If the text contains a large number of words that have less effect on the sentiment classification, it will reduce the classification accuracy of the emotional theme model. Based on this point of view, we should improve the influence of emotional words and emotional related words in the sampling process. The steps of the whole algorithm are shown in Fig. 3.

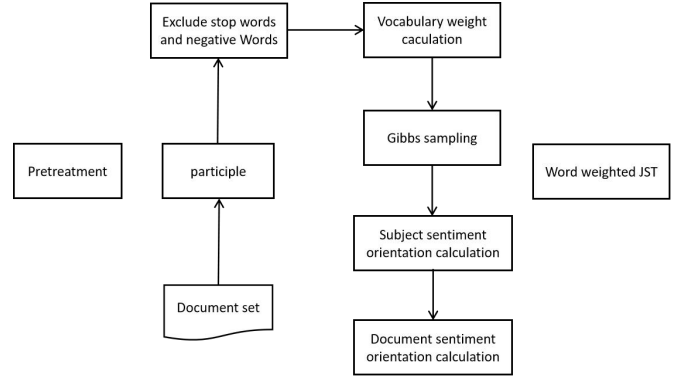


Fig. 3. Word weighted JST algorithm

This paper introduces three ways to incorporate word weights. One is to integrate the weight of emotional words into the vocabulary by calculating the point mutual information (PMI) of the words in the dictionary and the emotional seed words[12]. We use K-means method to obtain emotional seed words. For the word w_i , its weight formula is:

$$weight_{sent}(w_i) = \frac{1}{x} \sum_{e=1}^x \log \frac{p(w_i, positive(e))}{p(w_i)p(pos(e))} - \frac{1}{y} \sum_{f=1}^y \log \frac{p(w_i, negative(f))}{p(w)p(negative(f))} \quad (1)$$

In (1), x is the number of positive seed words, and y is the number of negative seed words.

Another weighting method is inspired by the literature[13], we introduce the feature weighting method of Gaussian function. For a news document, there may not be too many emotional words with obvious inclinations. This situation will lead to the frequency of emotional words appearing in the corpus is too low. However, the topic model represented by LDA or its improved word distribution will generally tilt to high frequency words, which will affect the accuracy of model emotion classification. In order to reduce the weight of high frequency words and increase the weight of the middle and low frequency words, we use the Gaussian function to weight the word i in the document:

$$weight_{gauss}(w_i) = \exp\left(-\frac{(f_i - f_{mid})^2}{2\sigma^2}\right)$$

In the formula, the variance σ^2 is:

$$\sigma^2 = \frac{1}{V-1} \sum_{i=1}^V (f_i - f_{mid})^2 \quad (2)$$

In (2), f_i is the word frequency of the word i , and f_{mid} is the median of the word frequency.

The last weighting method is to integrate the position information of the word. If the word appears in the headline, subtitle or summary part of the news, the word i is weighted:

$$weight_{title}(w_i) = \xi \cdot \ln(f + e - 1) \quad (3)$$

In the (3), ξ is the elastic coefficient, and $0 < \xi < 2$. The size of this parameter determines the influence of positional features on word weights. If we need to weaken the impact of location information, then ξ takes $0 < \xi \leq 1$. If we need to enhance the impact of location information, then ξ takes $1 < \xi \leq 2$. f is the word frequency of the word i .

The final word weight is calculated using (4).

$$\text{weight}(w_i) = \mu \cdot \text{weight}_{\text{sent}}(w_i) + \varepsilon \cdot \text{weight}_{\text{gauss}}(w_i) + \text{weight}_{\text{title}}(w_i) \quad (4)$$

In (4), μ, ε is a weighting parameter, $1 \leq \mu, \varepsilon \leq 2$, and $\mu + \varepsilon \leq 3$.

Using the above method that merge the weights of the words, each word can be treated in a targeted manner during the Gibbs sampling process.

In the JST algorithm, in order to obtain the distribution of π , θ and φ in the model, we first calculate the posterior distribution $p(z, l | w)$. That is, under the condition that the word w is given, we calculate the conditional probability that the subject of the word is z and the emotion label is l . We use the full probability formula to convert the conditional probability.

$$P(z, l | w) = P(z, l, w) / P(w) \quad (5)$$

From (5), we know $P(z, l | w) \propto P(w, z, l)$. So we only need to calculate the joint probability distribution of z, l, w . The joint probability distribution of z, l, w :

$$P(w, z, l) = P(w | z, l) \cdot P(z | l, d) \cdot P(l | d) \quad (6)$$

The $P(w | z, l)$, $P(z | l, d)$ and $P(l | d)$ in (6) correspond to the generation process of the document respectively. We calculate $P(w | z, l)$, $P(z | l, d)$ and $P(l | d)$ in (6). Then by integrating the weights of φ , θ , π and merging words, we can get:

$$P(w | z, l) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^{T \cdot S} \prod_j \prod_k \frac{\prod_i \Gamma(|\text{weight}(c)| * N_{j,k}^{(c)} + \beta)}{\Gamma((\sum_{c=1}^V |\text{weight}(c)| * N_{j,k}^{(c)} + V\beta))} \quad (7)$$

$$P(z | l, d) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \right)^{S \cdot D} \prod_k \prod_d \frac{\prod_j \Gamma((\sum_{i=1}^V |\text{weight}(i)| * N_{j,k,d}^{(i)} + \alpha)}{\Gamma((\sum_{j=1}^T (\sum_{i=1}^V |\text{weight}(i)| * N_{j,k,d}^{(i)} + T\alpha))} \quad (8)$$

$$P(l | d) = \left(\frac{\Gamma(S\gamma)}{\Gamma(\gamma)^S} \right)^D \prod_d \frac{\prod_k \Gamma((\sum_{m=1}^V |\text{weight}(m)| * N_{k,d}^{(m)} + \gamma)}{\Gamma((\sum_{k=1}^S (\sum_{m=1}^V |\text{weight}(m)| * N_{k,d}^{(m)} + S\gamma))} \quad (9)$$

In (7)(8)(9), V is the size of the index dictionary, and T is the total number of topics, S is the total number of emotional tags, and D is the total number of documents in the corpus. $N_{j,k}^{(c)}$ indicates the number of times that the word c is assigned to the subject j and the emotion tag k . $N_{j,k,d}^{(i)}$ represents the number of times that the word i in the document d is assigned to the emotion tag k and the subject j . $N_{k,d}^{(m)}$ represents the number of times that the word m in the document d is assigned to the emotion tag k . α is the asymmetric Dirichlet prior parameter of θ . β is the asymmetric Dirichlet prior parameter of φ . γ is the asymmetric Dirichlet prior parameter of π . Γ is the Gamma function.

Combining (6)(7)(8)(9), by using the properties of Gamma function and the conjugate property of Dirichlet, we can derive the conditional distribution probability of single sampling in the process of Gibbs sampling:

$$P(z_t = j, l_t = k | w, z_{-t}, l_{-t}, \alpha, \beta, \gamma, \text{weight}) = \frac{|\text{weight}(c)| * N_{j,k}^{(c-t)} + \beta}{(\sum_{i=1}^V |\text{weight}(c)| * N_{j,k}^{(c-t)} + V\beta)} * \frac{(\sum_{i=1}^V |\text{weight}(i)| * N_{j,k,d}^{(i-t)} + \alpha)}{(\sum_{i=1}^V |\text{weight}(i)| * N_{j,k,d}^{(i-t)} + T\alpha)} * \frac{(\sum_{m=1}^V |\text{weight}(m)| * N_{k,d}^{(m-t)} + \gamma)}{(\sum_{k=1}^S (\sum_{m=1}^V |\text{weight}(m)| * N_{k,d}^{(m-t)} + S\gamma))} \quad (10)$$

In (10), z_{-t} represents the subject distribution of all other words in the corpus except the t_{th} word in document d . l_{-t} indicates the emotional polarity of all other words in the corpus except the t_{th} word in document d .

We can use the Markov chain to estimate the topic-emotion-word distribution $\varphi_{c,j,k}$, topic-emotion-document distribution $\theta_{j,k,d}$ and emotion-document distribution $\pi_{k,d}$ respectively.

$$\varphi_{c,j,k} = \frac{|\text{weight}(c)| * N_{j,k}^{(c)} + \beta}{(\sum_{c=1}^V |\text{weight}(c)| * N_{j,k}^{(c)} + V\beta)} \quad (11)$$

$$\theta_{j,k,d} = \frac{(\sum_{i=1}^V |\text{weight}(i)| * N_{j,k,d}^{(i)} + \alpha)}{(\sum_{j=1}^T (\sum_{i=1}^V |\text{weight}(i)| * N_{j,k,d}^{(i)} + T\alpha))} \quad (12)$$

$$\pi_{k,d} = \frac{(\sum_{m=1}^V |\text{weight}(m)| * N_{k,d}^{(m)} + \gamma)}{(\sum_{k=1}^S (\sum_{m=1}^V |\text{weight}(m)| * N_{k,d}^{(m)} + S\gamma))} \quad (13)$$

In (11)(12)(13), α , β , γ are hyper-parameters, z is the subject of word, k is the emotional label of word, and $\text{weight}(l)$ is the weight of word.

The process of Gibbs sampling and document polarity determination in the word-weighted JST model are shown in Fig.4 below.

Input: Preprocessed news corpora
Output: The emotion of document
Specific steps:
1: Reading news texts, serializing words, building index dictionaries V ;
2: for $w \in V$:
3: Calculated $\text{weight}(w)$ according to (4);
4: Initialization matrix ϕ , θ , π ;
5: for $m=1$ to M
6: a) Read a word w_i from a document and randomly assign the topic tag to the word w_i .
7: while One word in the emotion dictionary
8: if $p = w_i$ do

```

9:      Assign the emotion tags of  $p$  to  $w_i$ ;
10:      end while;
11:      else
12:          Assign to  $W_i$  an emotional tag randomly;
13:      end while;
14:      b) Calculate the probability of the word  $w_i$  with subject  $j$  and emotion label
          $k$  according to (10).
15:      c) Reselect a topic  $j$  for the word based on the Markov chain
16:      d) Reselect an emotion tag  $k$  with subject  $j$ .
17:      e) Update matrix  $\phi, \theta, \pi$  according to new sampling results;
18:      f) Return to execution (a) Until all words are processed.
19: end for;
20: Calculate  $\pi_{k,d}$  according to (12);
21: if  $\pi_{k-,d} > \pi_{k+,d}$  do
22:     The emotion of document  $d$  is negative;
23: else if  $\pi_{k-,d} > \pi_{k+,d}$  do
24:     The emotion of document  $d$  is positive;
25: else do
26:     The emotion of document  $d$  is neutral.

```

Fig. 4. Gibbs sampling process and document polarity judgment in word-weighted JST model

IV. EXPERIMENTS AND RESULT

A. Experiment Setup

This paper randomly selects 2000 news, 745,892 words from the Vietnamese news corpus on the four themes of military, politics, technology and real estate. The size of the index dictionary V is 34,948. In order to facilitate the comparison of experimental results, we use the manual annotation method to label the positive and negative sentiment of these 2000 news. Among them, there are 880 positive emotional news and 610 negative emotional news and 510 neutral emotional news. This article uses skip-gram to train the word embedding model and the word vector dimension is 50 dimensions.

In order to verify the feasibility of this method, we set up three sets of comparative experiments:

In order to verify the feasibility of the proposed method in the paper, we set up three sets of comparative experiments:

In order to verify the validity of the merging word weighting method proposed in this paper, we set the weight comparison of some words in the three weighting methods in Experiment 1.

In order to verify the impact of the inclusion of word weight information on Vietnamese sentiment classification, we set up the second set of contrast experiments. In Experiment 2, we verified the influence of the integration of word weight information and non-integration word weight information on Vietnamese sentiment classification without incorporating the Vietnamese emotional dictionary.

In order to verify the impact of prior knowledge on Vietnamese sentiment classification, we set up the third set of comparative experiments. In Experiment 3, we verified the influence of the integration of emotional dictionary and non-integration sentiment dictionary on Vietnamese sentiment classification under the condition of merging word weights.

In the three groups of experiments, the model parameters α, β, γ were set to $50/T, 0.01$, and 0.01 respectively. We will set elastic coefficient ξ to 0.618 and set the weighting parameters μ, ε to 1.5, 1.2 respectively. Due to limited system resources, we have set the number of iterations to 2000 based on experience. In the experiment, the total number of different topics T has a great influence on the accuracy of the model classification, because the total number of topics set manually affects the model parameters. In the experiment, we will also display the number of topics T as an independent variable in the chart.

Finally, if a news is labeled with the same emotions as the manual, then we think the model correctly identifies the news document. Next, we will give a formula to evaluate the accuracy of the model:

B. Experimental Results and Analysis

In Experiment 1, we calculate the weight value of the sentiment word, the weight value of the Gaussian weight, the weight value of the position information, and the weight value of the fusion word proposed in this paper. The experimental results are shown in Table II.

Table II first shows the weighting values of some words in the way of the merging word weighting proposed in this paper. Then Table II shows the weight values of some words in the three word weighting modes. If the weight value of a word is larger, the distinction of the word is more in the process of sampling. It can be clearly seen from Table II that the first five words have obvious emotional tendencies, while the latter five words have more subject matter. In Table II, the “Senti” weighting method distinguishes the sentiment words very well, but it has a great influence on the topic distribution. The “Gaussian” method does not distinguish between emotional words and subject words. The “Title” method is to balance the weights of emotional words and subject words. Therefore, the way of merging words weights makes the emotional words more discriminating without seriously affecting the classification of the topics. The data in Table II shows that the way of merging words weights proposed in this paper is effective.

TABLE II. THE WEIGHT OF WORDS IN DIFFERENT WAYS

words	Weighted method			
	Senti+Gaussian+Title	Senti	Gaussian	Title
Lo lắng	8.43	3.39	1.31	1.78
Kiên quyết	8.12	3.26	0.925	2.12
thiệt hại	9.05	3.67	1.38	1.88
tuyệt giao	8.33	3.13	1.85	1.42
chửi mắng	7.88	3.36	1.425	1.13
Hoa Kỳ	3.78	0.53	0.71	2.14
Zod.	3.63	0.49	0.54	2.25
Hàn Quốc	4.21	0.46	1.075	2.23
sắp xếp	3.61	0.43	0.62	2.22
trang bị	1.55	0.33	0.488	0.47

In Experiment 2, we constructed the JST model and the word weighted JST model separately, and calculated the accuracy of the sentiment classification of Vietnamese news texts under different the number of topics.

It can be seen from Fig.5 that under the condition of no prior

knowledge, incorporating the weight of words into the model does improve the accuracy of sentiment classification. However, compared with the case of unweighted model, the weighted model is slightly more affected by the number of topics T than the former, that is, the stability is not as good as the former. The reason may be that the JST model is integrated with the weight of words, which increases the distance between words and words. This affects the process of word sampling iterations, which ultimately leads to a decrease in the stability of the model.

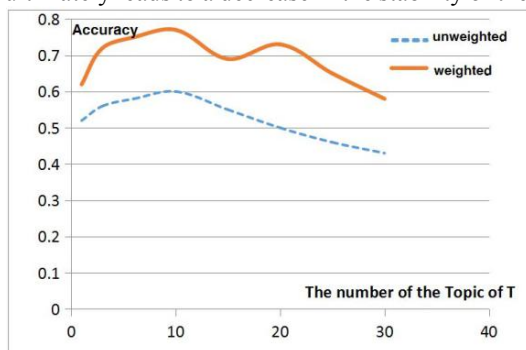


Fig. 5. Word Weighted and Unweighted Results Graphs Without Prior Knowledge

In Experiment 3, we constructed a word-weighted JST model that incorporates Chinese and Vietnamese emotional dictionaries and a word-weighted JST model that does not fit into Chinese and Vietnamese emotional dictionaries. Then we calculate the accuracy of the sentiment classification of Vietnamese news texts under different number of topics. The experimental results are shown in Fig.5.

As can be seen from Fig.6, prior knowledge can improve the accuracy of the sentiment classification of the model to some extent. As can be seen from the results graph, after the number of topics is 20, the effect of prior knowledge on the results is not very large. When the number of topics T is between 5 and 12, the word-weighted JST model has the best sentiment classification effect.

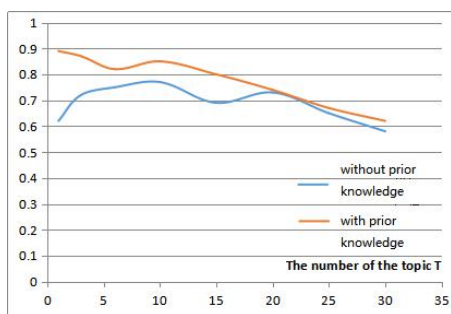


Fig. 6. Influence of word weights with/without prior knowledge on classification results

V. CONCLUSIONS

This paper proposes an unsupervised classification model of news in Chinese and Vietnamese: word weighted JST model. In order to solve the problem that the JST model randomly assigns emotional tags during the sampling process, this paper proposes to use the sentiment dictionary to label the emotional tags of the words. Then, in order to solve the problem of less discrimination between emotions and topics in the JST model, the model proposes to assign different weights to different words in Gibbs sampling. The final experiment shows that the word-weighted JST model combined with prior knowledge has greatly improved the accuracy of Vietnamese sentiment classification. The next step will be to focus on how to incorporate emotional information at the sentence level.

REFERENCES

- [1] X. JieZhou, W.XiaoJun, X.Jianguo, "Attention-based LSTM Network for Cross-Lingual Sentiment Classification", Proceedings of EMNLP, 2016,247-256
- [2] Z.Huiwei, Y.Yunlong, L.Zhuang, "Jointly Learning Bilingual Sentiment and Semantic Representations for Cross-Language Sentiment Classification", China Conference on Information Retrieval, 2017,149-160
- [3] C.Yongheng, Z.Wanli, L.Yaojin. "Emotional analysis method based on topic seed words", computer application, 2015, 35(9):2560-2564.
- [4] OY.Jihong, I.Yanhui, L Ximing,"Multi-granularity theme emotion blending model based on LDA".Electronic Journal, 2015, 43(9):1875-1880.
- [5] Z.Jianhua, L.zhengyou, "A Sentiment Analysis Method Based on Sentiment Words Extraction and LDA Feature Representation", Computer and Modernization, 2014(5) :79-83
- [6] Lin C, He Y, "Joint sentiment/topic model for sentiment analysis", ACM Conference on Information and Knowledge Management, 2009:375-384.
- [7] P.Yunxian, Y.Fang, "Affective Classification of News Text Based on JST Model", Zhengzhou University Journal of Science, 2015, 47(1):64-68.
- [8] K.Wang, X.Rui, "A Summary of Automatic Construction Methods of Emotional Dictionary", Acta Automatica Sinica, 2016, 42(4):495- 511.
- [9] L.RongJun, W.Xiaojie, Zhou.Yanquan, "Application of PageRank Model in Chinese Emotional Word Polarity Discrimination". Journal of Beijing University of Posts and Telecommunications, 2010, 33(5):141-144.
- [10] Z.Chengong, L.Peiyu, Z.Zhenfang, "A sentiment analysis method based on polarity dictionary", Journal of Shandong University (Science Edition), 2012, 47(3):50-53.
- [11] Z.yongmei, Y.jianeng, Y.aimin. "A method on building Chinese sentiment lexicon for text sentiment analysis", Journal of Shandong University, 2013(6) :27-33
- [12] Turney P D, Littman M L, "Measuring praise and criticism:Inference of semantic orientation from association", Acm Transactions on Information Systems, 2003, 21(4):315-346.
- [13] Z.Xiaoping. "Thematic model and its application in clinical diagnosis and treatment of traditional Chinese medicine". Beijing Jiaotong University, 2011.