2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Robust Framework for Short Text Categorization based on Topic Model and Integrated Classifier

Peng Wang, Heng Zhang,Yu-Fang Wu, Bo Xu and Hong-Wei Hao
Institute of Automation of the Chinese Academy of Sciences
Beijing, 100190 P. R. China
peng.wang@ia.ac.cn, hzhang07@nlpr.ia.ac.cn

***Abstract*—In this paper, we propose a method for short text categorization using topic model and integrated classifier. To enrich the representation of short text, the Latent Dirichlet Allocation (LDA) model is used to extract latent topic information. While for classification, we combine two classifiers for achieving high reliability. Particularly, we train LDA models with variable number of topics using the Wikipedia corpus as external knowledge base, and extend labeled Web snippets by potential topics extracted by LDA. Then, the enriched representation of snippets are used to learn Maximum Entropy (MaxEnt) and support vector machine (SVM) classifiers separately. Finally, viewing that the most possible predicted result will appear in the top two candidates selected by MaxEnt classifier, we develop a novel scheme that if the gap between these candidates is large enough, the predicted result is considered to be reliable; otherwise, the SVM classifier will be integrated with MaxEnt classifier to make a comprehensive prediction. Experimental results show that our framework is effective and can outperform the state-of-the-art techniques.**

I. INTRODUCTION

With the vigorous development of web 2.0, the short text have flooded everywhere, including micro-blog, products review, short messages, advertising messages and search snippets. Therefore, the high-precision classification of short text will make significant assistance on these web applications. Sun A. [1] proposed a simple method for short text categorization by selecting the most representative words as query to search a few of labeled samples, and the majority vote of the search results is the predictable category.

In most traditional text mining tasks, documents are represented based on the statistical methods such as bag of words model [2], which ignores the textual information and lacks semantic knowledge. Especially for short text with small length, one encounters serious data sparsity problem in presentation. For instance, if two short texts use different sets of key words to cover the same topic, it is difficult to measure their similarity [3]. Most of current popular techniques to solve this problem are to extend short text representation using latent semantics or related words. In the current research, the enriching information may be derived using topic model internally from the short text collection [4], or from a larger external knowledge base such as Wikipedia [5]-[8] or using knowledge graph like WordNet [9].

Obviously, in order to enhance the short text classification performance, we should enrich the short text without introducing too much noise to the greatest extent. Wikipedia is the most influential thesaurus on the web with millions of well-formed articles [4]. Gabrilovich E. and Markovitch S. [10] proposed a method to improve text classification performance by enriching document representation with Wikipedia concepts. In this paper, using Wikipedia as external corpus, we explore the latent semantics based on LDA which is a generative graphical model, and widely used in text mining [11]. Then we can expand the short text by appending semantics to it, with the aim of enriching representation and reducing sparseness of short text.

After the feature extraction procedure of text, how to utilize the previous results to learn a satisfied classifier is also an inevitable problem. Among various machine learning methods, MaxEnt and SVM have been successfully applied in many text mining tasks [12], which proves that MaxEnt is much faster in both training and inference while SVM is more robust. In this paper, we study the problem of how to integrate them to fully explore their advantages. Firstly, before classifying each test sample exactly, we select these classes with larger score as candidates by MaxEnt classifier. Then we make a decision that if the gap between these candidates is large enough, the predictable result might be reliable; otherwise, the SVM classifier will be integrated with MaxEnt classifier to make a comprehensive prediction. Based on an open database [6], extensive experiments demonstrate the effectiveness of our framework, which have outperformed the state-of-the-art methods.

The rest of this paper is organized as follows. We firstly review relevant works in section II. Then we simply introduce the theoretical background of topic model and propose our classification framework in section III. We systematically validate the method over experiments in section IV, followed by the conclusion in section V.

II. RELATED WORK

In past decades, short text classification has been an active research area and has drawn lots of attention. To our knowledge, the general methods for normal text mining cannot apply to short text tasks directly, because of data sparsity and noise. Hence, some highly related text segments may have very little overlapping on the word level which

This work is supported by the National Natural Science Foundation of China (NSFC) Grant No.61203281.

poses great challenges for similarity measurement [3] and impacts the categorization performance seriously. In this section, we review existing techniques for categorization of short text that are related to our framework.

In recent years, there are some researches on how to utilize large-scale data collection to explore semantics which can be used to enrich features derived from bag-of-words representation, and help text understanding [4]-[7],[13]. Empirical evaluation shows that by resolving synonyms and introducing concepts, the semantics can improve the quality of document categorization.

Modeling short text based on Wikipedia and LDA by Phan et al. [6] is probably the most relevant work to our study. Phan X.H. proposed the method for using Wikipedia as external corpus to train LDA model which discover hidden topics by Gibbs sampling, and appending topic names to original text to resolve the data sparseness problem in short text classification. Although the approach is proved to be enhanced, it does not fully utilize the inference result of LDA to model a more powerful classifier. Zhu Y. et.al. [8] assume that single external dataset may not cover enough topics so they propose to use multi-original external corpus and adjust the weight of features captured from short texts to identify topics for better categorization performance. Compared to identifying topics with LDA directly based on one external corpus, the topics produced by this approach are more broad and accurate. However, these researches mainly focus on extracting topics of certain granularity which are usually not sufficient to set up effective feature space. Viewing this reason, Chen M. et.al. [5] pointed out that leveraging topics at multiple granularity can model short texts more precisely.

Previous methods, which suffer from severe data sparsity of short text, typically rely on the document-level word co-occurrence to reveal topic structure. Yan X. [14] proposed a method for modeling topics over the whole corpus instead of document-level to cope with the problem of sparsity. Unlike probabilistic formulation of topic model above, Zhu J. and Xing P. [15] present a non-probabilistic one named sparse topical coding, which can control the sparsity of inferred result directly.

III. METHODOLOGY

A. Topic Model

With respect to topic modeling, Latent Dirichlet Allocation (LDA) [11] is a method to explore the latent semantic structure, which is to perform latent semantic analysis (LSA) [16], which can be recovered by the co-occurrence of terms in documents. Obviously, LDA is closely related to probabilistic Latent Semantic Analysis (pLSA) proposed by Hofmann [17], a probabilistic formulation of LSA. While it is widely acknowledged that LDA has more complete foundation than pLSA in that it follows a full probabilistic generation process for text collection [18].

Blei et.al. [11] firstly proposed LDA and used it to estimate the multinomial observations by unsupervised learning. As depicted in both Fig. 1 and Table I, LDA was developed

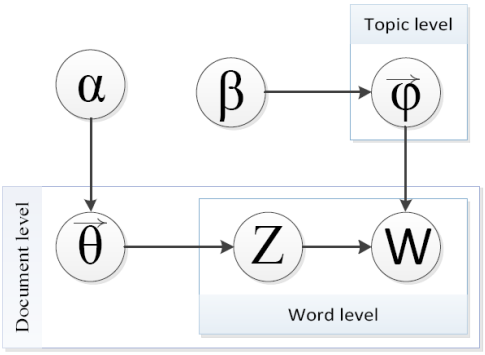

Fig. 1. Graphical model of LDA

based on an assumption of document generation process, which can be interpreted as follows. For a document $\overrightarrow{D_m} = \{w_{m,n}\}$, where $m = 1,2,\cdots,M$, $n = 1,2,\cdots,N_m$, multinomial distribution $\overrightarrow{\theta_m}$ over topics is firstly sampled from a Dirichlet distribution $Dir(\alpha)$, which determines topic assignment for words in that document. Before the word $w_{m.n}$ being generated, a particular topic name $Z_{m,n}$ is extracted from $\overrightarrow{\theta_m}$, which perform the topic assignment. Then the word $w_{m.n}$ is generated by sampling from multinomial distribution $\overrightarrow{\varphi_{Z_{m,n}}}$, which is drawn from another Dirichlet distribution $Dir(\beta)$.

According to both the simple Bayesian graphical model depicted in Fig. 1 and the description of document generation process above, we can write the joint distribution of all observations and hidden variables as follows.

$$\begin{aligned} &p(\overrightarrow{D_m}, \overrightarrow{Z_m}, \overrightarrow{\theta_m}, \mathbf{\Phi} \mid \alpha, \beta) \\ &= p(\mathbf{\Phi} \mid \beta) \prod_{n=1}^{N_m} p(w_{m,n} \mid \overrightarrow{\varphi_{Z_{m,n}}}) p(Z_{m,n} \mid \overrightarrow{\theta_m}) p(\overrightarrow{\theta_m} \mid \alpha) \end{aligned} \tag{1}$$

Then to integrate over $\overrightarrow{\theta_m}$, $\mathbf{\Phi}$ and sum over $\overrightarrow{Z_m}$ based on formula (1), viewing that $w_{m,n}$ is independent of $Z_{m,n}$ when given $\overrightarrow{\theta_m}$ and $\mathbf{\Phi}$, we can obtain the likelihood of document $\overrightarrow{D_m}$ as

$$\begin{aligned} &p(\overrightarrow{D_m} \mid \alpha, \beta) \\ &= \iint p(\mathbf{\Phi} \mid \beta) p(\overrightarrow{\theta_m} \mid \alpha) \cdot \\ &\qquad \sum_{\overrightarrow{Z_m}} \prod_{n=1}^{N_m} p(w_{m,n} \mid \mathbf{\Phi}, Z_{m,n}) p(Z_{m,n} \mid \overrightarrow{\theta_m}) d\mathbf{\Phi} d\overrightarrow{\theta_m} \\ &= \iint p(\mathbf{\Phi} \mid \beta) p(\overrightarrow{\theta_m} \mid \alpha) \sum_{\overrightarrow{Z_m}} \prod_{n=1}^{N_m} p(w_{m,n}, Z_{m,n} \mid \mathbf{\Phi}, \overrightarrow{\theta_m}) d\mathbf{\Phi} d\overrightarrow{\theta_m} \\ &= \iint p(\mathbf{\Phi} \mid \beta) p(\overrightarrow{\theta_m} \mid \alpha) \prod_{n=1}^{N_m} p(w_{m,n} \mid \mathbf{\Phi}, \overrightarrow{\theta_m}) d\mathbf{\Phi} d\overrightarrow{\theta_m} \quad . \end{aligned} \tag{2}$$

Finally, the likelihood of the whole corpus $\mathbf{D} = \{\overrightarrow{D_m}\}$ is inner-product of the likelihoods of all documents as in (3) when the Dirichlet parameters are given.

TABLE I. PARAMETERS AND VARIABLES USED IN LDA MODEL

Parameters	Details
M	the number of documents in corpus
N_m	the length of mth document
α, β	parameters for Dirichlet distribution
$\overrightarrow{\theta_m}$	topic distribution in document m
$w_{m,n}$	a particular word for word placeholder $[m,n]$
$Z_{m,n}$	topic index of nth word in document m
$\overrightarrow{\varphi_{Z_{m,n}}}$	word distribution for topic $Z_{m,n}$
V	vocabulary size
K	the number of topics
$\mathbf{\Theta} = \{\overrightarrow{\theta_m}\}$	a $M \times K$ matrix
$\mathbf{\Phi} = \{\overrightarrow{\varphi_{Z_{m,n}}}\}$	a $K \times V$ matrix
$\mathbf{D} = \{\overrightarrow{D_m}\}$	corpus with M documents

$$p(\mathbf{D} \mid \alpha, \beta) = \prod_{m=1}^{M} p(\overrightarrow{D_m} \mid \alpha, \beta) \tag{3}$$

As demonstrated in Table II that the words from each topic are very related, LDA make the presentation of short text more topic-focused, which allows to model linguistic phenomena like synonymy and polysemy to alleviate the data sparsity and noise and perform dimensionality reduction to some extent. With increasingly attention paid to topic model applications in various text mining communities, LDA is becoming a standard technique in topic modeling. At the same time, a number of extensions and variants to LDA have been proposed.

B. Machine learning methods

In this paper, we choose MaxEnt [12] and SVM to build the integrated framework and as baseline classifiers to make comparisons with experiments in [5], [6]. There has been many successful applications of them in various Natural Language Processing tasks for its superior properties.

In general, according to MaxEnt principle, one can deduce a log-linear model which has the following form,

$$p(c_i \mid \overrightarrow{D_m}) = \frac{1}{pf(\overrightarrow{D_m})} \exp(\sum_j \lambda_j f_j(\overrightarrow{D_m}, c_i)) , \tag{4}$$

where c_i indicate the ith class, and λ_j is the weight of jth feature, which measures the contribution of jth feature to the model.

TABLE II. MOST LIKELY WORDS OF SOME TOPICS FROM WIKIPEDIA

Topic0:music band rock album song songs released bands records
Topic1:species food animals animal plants humans fish plant birds
Topic2: energy mass field quantum particles force theory system
Topic3: india indian hindu pakistan sanskrit century buddhist
Topic4: blood body brain heart cells muscle syndrome pressure
Topic5: water carbon oil chemical gas process oxygen acid
Topic6: government party president constitution election minister
Topic7: power energy solar electric electrical voltage circuit
Topic8: ystem data code software computer program systems
Topic9: horse opponent horses body hand match foot wrestler
Topic10: south africa united country islands world african spanish

$$pf(\overrightarrow{D_m}) = \sum_i \exp(\sum_j \lambda_j f_j(\overrightarrow{D_m}, c_i)) , \tag{5}$$

Where $pf(\overrightarrow{D_m})$ is a normalization factor to ensure that (4) satisfy probability constraint condition. MaxEnt model represents features with binary functions known as in (6), which maps a pair of feature and label of category to {1, 0}.

$$f_{w,c'}(\overrightarrow{D_m}, c_i) = \begin{cases} 1 , & c_i = c' \text{ and } w \in \overrightarrow{D_m} \\ 0 , & \text{otherwise} \end{cases} , \tag{6}$$

where w is a word from vocabulary. The intuition behind this feature function indicate that the outcome of maps for short text is very sparse, so the training and inference processes of MaxEnt are more rapid and effective than other techniques.

Additionally, SVM classification algorithm [19] used to solve two-class problems, are based on finding hyper-planes with maximal margin, which are defined by support vectors as in (7).

$$\begin{aligned} f(\overrightarrow{D_m}) &= sign(\sum_{i=1} \lambda_i y_i \mathbf{K}(\overrightarrow{D_m}, \overrightarrow{D_i}) + b) \\ &= sign(\sum_{i=1} \lambda_i y_i \phi^T(\overrightarrow{D_i}) \phi(\overrightarrow{D_m}) + b) \\ &= sign(\mathbf{w}^T \phi(\overrightarrow{D_m}) + b) , \end{aligned} \tag{7}$$

Where $\mathbf{K}(\overrightarrow{D_m}, \overrightarrow{D_i}) = \phi^T(\overrightarrow{D_i}) \phi(\overrightarrow{D_m})$ is the positive definite kernel function, λ_i is the Lagrange multiplier in dual problem corresponding to i th support vector, and $\phi(\bullet)$ is a mapping function especially used for non-linear classification problems.

Because the goal of SVM is to measure the margin of separation of the feature points instead of matching on them, this character makes SVM can handle even fairly large feature sets well. While the representation of short text usually accompanied by high dimensional feature space since each stemmed word is a feature, so SVMs scale well and have good performance in document classification.

C. Our framework

The overall framework of our method based on LDA and integrated classifier to practically improve short text classification is presented in Fig. 2. We develop this method based on the work of professor phan [6] that the LDA model is estimated using large-scale Wikipedia corpus as external knowledge-base, then perform inference over Web snippets. While in this paper for estimating LDA model or conducting inference with variable number of topics instead of single topic which is one of the differences from [6]. Another improvement is that we learn different types of categorization models using enriched representation of training samples and build integrated classifier based on them.

The particular procedure of our framework is as follows. Firstly, we train topic models leveraging Wikipedia corpus and extract latent information by topic inference. Secondly, the representation of training and test dataset is expanded using hidden topics with different granularity. The way of feature expansion is to combine the bag of words feature and topic feature from the topic space derived from Wikipedia.

TABLE III. DETAILS OF WEB SNIPPETS

Domain	Number of training snippets	Number of test snippets
Business	1200	300
Computers	1200	300
culture-arts-entertainment	1880	330
Education-Science	2360	300
Engineering	220	150
Health	880	300
Politics-Society	1200	300
Sports	1120	300
Total	10060	2280

Thirdly, MaxEnt and SVM are learned by the enriched features with different topic number respectively. Finally, we build the integrated classifier under the principles as follows. MaxEnt is chosen as the initial classifier to select top two candidates for most possibly predicted classes. Then the integrated classifier works according to the decision rule that if the gap between candidates meet the threshold obtained by cross validation previously, the initial classification is viewed as a reliable prediction, otherwise SVM will be integrated to make a comprehensive prediction with (8) .

$$Score = \alpha * p(c_i \mid \overrightarrow{D_m}) + (1-\alpha) * \log(f(\overrightarrow{D_m})) \ , \qquad (8)$$

Where α is a weight of integration, which is obtained by cross validation. As described above, our framework is easy to be implemented and proved to be robust with comparable running time to single classifier.

The highlighted advantages of our framework lie in that the enhanced representation of short text with different topics provide multi-granularity discriminant features substantially and the global classifier MaxEnt integrates local classifier SVMs together which make them supplement each other to ensure the efficiency and precision of our framework. Additionally, SVM is especially used to handle a two-class categorization task which can be fully exploited in our framework.

IV. EXPERIMENTS

Based on Wikipedia corpus and Web snippets dataset that has been used in [5], [6], We carry out experiments to evaluate our method and make comparisons with them.

A. Experimental data

With similar experimental settings to [5], [6], we use Wikipedia corpus as external large-scale knowledgebase to train LDA model, which contains 71986 documents and is crawled by Phan X. H. Additionally, Web snippets dataset collected also by Phan X. H., consists of 10,060 training snippets and 2,280 test snippets from 8 categories, shown in Table III. On average, each snippet has 18.07 words.

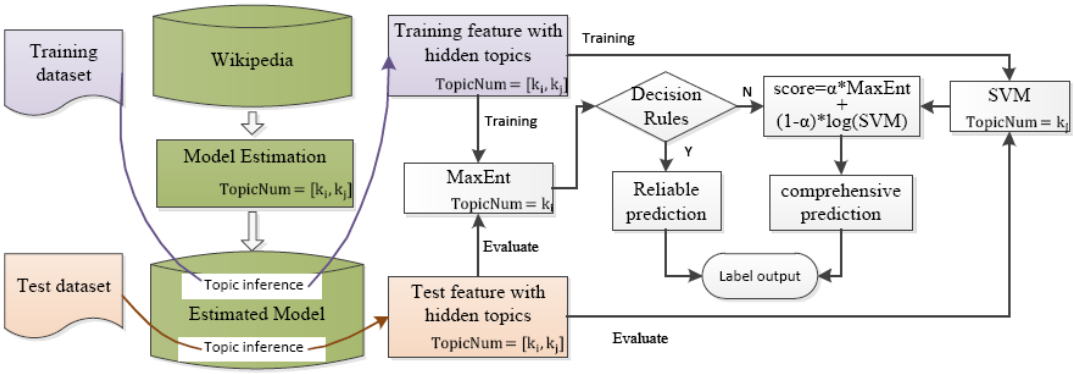

Fig. 2. The framework based on topic model and integrated classifier

TABLE IV. Accuracy of integrated classifier vary with k_i and k_j

k_j \ k_i	10	20	30	40	50	60	70	80	90	100	150
10	74.56	85.18	84.25	84.17	85.53	85.35	85.44	82.72	82.54	83.55	77.68
20	84.6	83.64	85.96	86.05	86.97	86.27	86.55	84.56	84.74	85.48	82.72
30	84.3	85.04	82.72	85.26	86.71	86.27	85.83	83.95	84.43	85.22	82.11
40	84.43	85.35	85.2	82.59	85.96	85.83	84.3	83.99	84.34	84.91	82.24
50	85.39	86.32	86.23	85.88	83.51	85.92	86.27	84.12	84.12	85.22	82.46
60	85.04	86.36	85.61	85.57	86.1	82.81	84.96	84.12	84.43	85.18	82.68
70	84.74	85.92	85.18	84.78	86.27	84.73	83.07	83.51	83.9	84.96	82.32
80	83.07	84.78	84.34	84.47	84.87	84.3	84.43	79.96	83.33	83.68	80.04
90	84.08	85.22	84.91	84.91	85.92	84.78	84.91	83.46	80.70	83.46	81.71
100	83.42	85.09	85.09	85.66	86.32	85.0	85.09	83.77	83.42	82.5	81.36
150	82.89	84.96	85.0	85.26	85.92	85.35	85.31	82.89	83.77	83.86	78.51

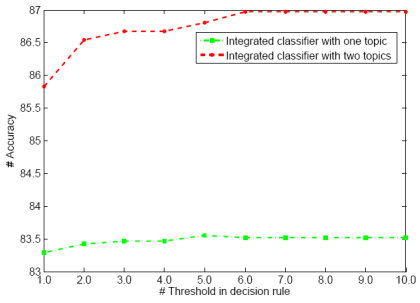

Fig. 3. Accuracy of classifiers over ε

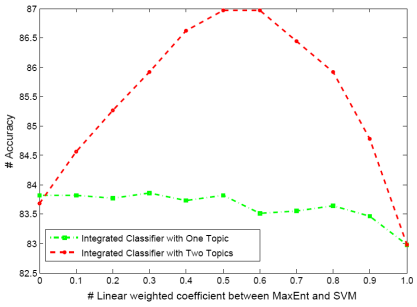

Fig. 4. Accuracy of classifiers over α

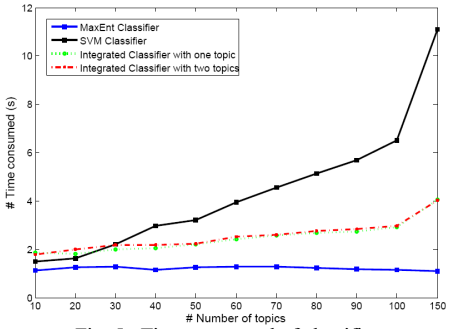

Fig. 5. Time consumed of classifiers over variational topics

B. *Experimental results and analysis*

In order to distinctly reveal the effectiveness of our framework as depicted in Fig. 2, firstly we estimate different topic models with topic number $k = \{10, 20, \cdots, 100, 150\}$. Before obtaining the integrated classifier, we select LDA models with k_i, k_j topics to extract topic names to perform feature expansion. Then use these enriched features to train MaxEnt classifier and respectively. At last, if $k_i = k_j$, we obtain an integrated classifier with single topic; otherwise with two topics. We carry out five-fold cross validation experiments and find that nearly when $\alpha = 0.6$, and the threshold $\varepsilon = 0.8$ in the decision rule, the integrated classifier can obtain more favorable results which is demonstrated in both Fig. 3 and Fig. 4. These results also imply that the MaxEnt classifier plays slightly more important role than SVM classifier in our framework.

With the setting of $\alpha = 0.6$, $\varepsilon = 0.8$, the experimental results are showed in table IV. From table IV, we can find that the accuracy of integrated classifier vary significantly with the changes of topic number k_i, k_j and the scenes of two topics generally have higher accuracy than one topic which also have been proved by [5]. while we obtain the highest accuracy of 86.97% when $k_i = 50$ and $k_j = 20$ which reduce classification error by 11.3% compared to [5] and by 26.88% compared to [6].

Taking the results of table IV into consideration, we select $k_i = 50$, $k_j = 50$ or 20 to build the special one-topic or two-topics integrated classifier. With $\alpha = 0.6$, ε ranging from 1 to 10, and keeping $\varepsilon = 8.0$, changing α from 0 to 1.0, we carry out experiments, then the results are illustrated in Fig. 3 and Fig. 4 separately. Fig. 3 mainly highlights two points: firstly, with the ranges of ε, the integrated classifier employing two topics always outperform the single topic one. Secondly, the performances of integrated classifiers become stable when $\varepsilon \geq 6.0$. Fig. 4 demonstrates that different from single topic cases, α makes significant influence on two-topics integrated classifier and our framework achieves highest improvement when $\alpha = 0.5$ or 0.6.

Furthermore, with the purpose of making overall comparisons, MaxEnt classifier and SVM classifier are used as baselines in the following experiments. For two-topics integrated classifier, maintaining $k_i = 50$, k_j varies from 10 to 100,150. When $\alpha = 0.6$, $\varepsilon = 8.0$, experimental results are respectively demonstrated in Fig. 6 and Fig. 7, which consistently indicate that our approach always outperform baselines definitely except the case of $k_j = 50$.

Since the integrated classifier employed two topics degenerate into one-topic scene when $k_j = 50$, its performance decrease obviously. Moreover, the one-topic integrated classifier doesn't apparently manifest superiority over the changes of k_j compared to baselines, which once again proves that multi-granularity topics can provide more discriminate information than single topic. It is also notable in Fig. 7 that our method obtains the best micro-average precision of 88.42% and the best micro-average recall of 86.6%.

At last, the time consumed of classifiers over variational topics in prediction is compared in Fig. 5. We can find that the MaxEnt classifier is faster and more stable than others. The SVM classifier will consume more time with the increase of the dimensionality of features and dominate the efficiency of our framework. Our integrated classifiers work according to the decision rule described in Section C of part III, which can accelerate the predicting procedure, so the time consumed of our method is between SVM and MaxEnt classifiers.

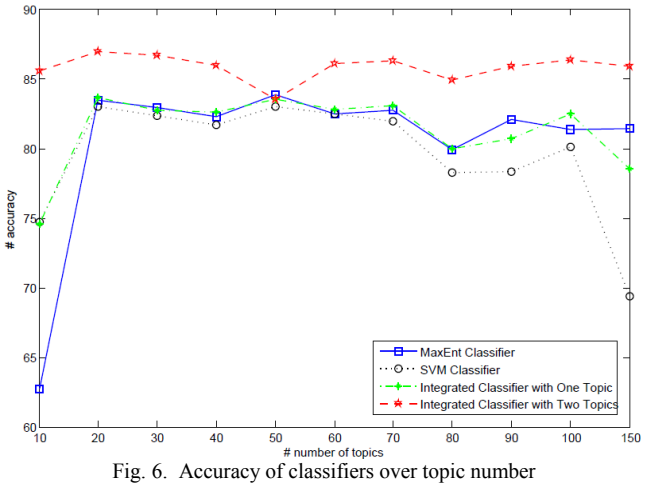

Fig. 6. Accuracy of classifiers over topic number

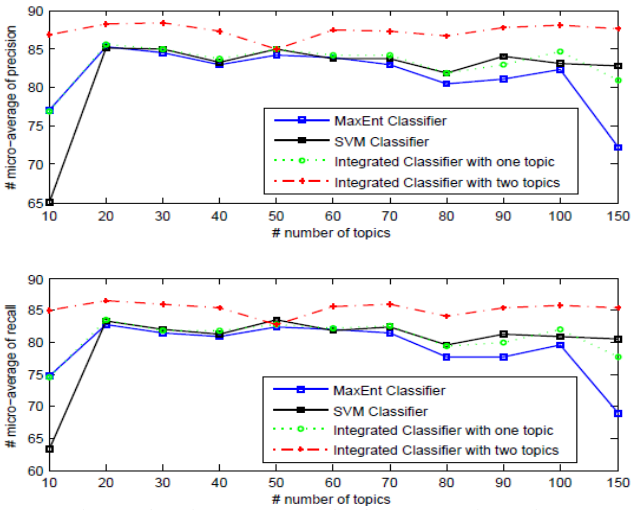

Fig. 7. The micro-average evaluations over topic number

V. Conclusion and Future Work

Due to the short length of the text segments, they don't provide enough discriminative features, and most of the normal text mining algorithms can't be employed directly in the short text analytical tasks with their full potentials. Additionally, semantics of the context is seriously ignored in the bag of words model which result in notorious sparseness. Topic model can be used to identify the latent information to expand the representation of short text [11].

In this paper, we present a robust framework for leveraging Wikipedia as external corpus to extract semantics via LDA to enrich the presentation of short text and integrate MaxEnt classifier and SVM classifier to improve text categorization performance. In our framework, MaxEnt is firstly used to select candidates of target class. Then if decision rule is met, we obtain a reliable prediction; otherwise MaxEnt integrate SVM to give a comprehensive prediction which leads to a faster training process and better quality. To demonstrate the effectiveness of our approach, based on an open dataset we conduct experiments and make systematic comparisons with baselines. The results mainly show three points as follows. Firstly, enriching document representation using latent topics extracted by LDA can validly reduce the data sparseness and achieve significant quality enhancement in the classification performance compared to the current techniques. Secondly, the linear weighted coefficient between classifiers and the threshold in decision rule can make difference to the performance of our framework. Thirdly, multi-granularity topics would enhance the presentation of document, so the integrated classifier leveraging two topics outperform the single topic scene.

However, the augmented representation of short text using latent topics may introduce noisy, so we will further study the scheme of how to effectively use the inference results of LDA to expand documents in text mining tasks. Furthermore, how to draw more meaningful topic structure as supplementary feature in classification problems by substantially different topic models and build our special external large corpus should be taken into consideration in the future. We believe that the proposed approach has great potential to achieve much better results after resolving these problems described above.

Acknowledgment

We thank professor Phan X. H. for sharing experimental data and his implementation of LDA.

References

[1] Sun A. "Short text classification using very few words", Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 1145-1146, 2012.

[2] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. "Short text classification in twitter to improve information filtering", Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 841-842, 2010.

[3] Zhang, L., Li, C., Liu, J., Wang, H. "Graph-based text similarity measurement by exploiting Wikipedia as background knowledge". World Academy of Science, Engineering and Technology, vol. 59, pp. 1548-1553, 2011.

[4] Hu X., Sun N., Zhang C., and Chua T. S. "Exploiting internal and external semantics for the clustering of short texts using world knowledge", Proceedings of the 18th ACM conference on Information and knowledge management. ACM, pp. 919-928, 2009.

[5] Chen M., Jin X., Shen D. "Short text classification improved by learning multi-granularity topics", Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume. AAAI Press, vol. 3, pp. 1776-1781, 2011.

[6] Phan X. H., Nguyen L. M., Horiguchi S. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections", Proceedings of the 17th international conference on World Wide Web. ACM, pp. 91-100, 2008.

[7] Hu X., Zhang X., Lu C., Park E. K., and Zhou X. "Exploiting Wikipedia as external knowledge for document clustering", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 389-396, 2009.

[8] Zhu Y., Li L., and Luo L. "Learning to classify short text with topic model and external knowledge", Knowledge Science, Engineering and Management. Springer Berlin Heidelberg, pp. 493-503, 2013.

[9] H. Andreas, S. Staab, and G. Stumme. "Ontologies improve text document clustering." Data Mining, pp. 541–544, 2003.

[10] Gabrilovich Evgeniy, and Shaul Markovitch. "Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge." AAAI. vol. 6, pp. 1301-1306, 2006.

[11] Blei D. M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the journal of machine learning research, vol. 3, pp. 993-1022, 2003.

[12] Berger Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. "A maximum entropy approach to natural language processing." Computational linguistics, vol. 22.1, pp. 39-71, 1996.

[13] E. Gabrilovich and S. Markovitch. "Computing semantic relatedness using Wikipedia-based explicit semantic analysis". In IJCAI, vol. 7, pp. 1606-1611 2007.

[14] Yan X. H., Guo J. F., Lan Y. Y, and Cheng X. Q. "A biterm topic model for short texts", Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 1445-1456.

[15] Zhu J., Xing E. P. "Sparse topical coding", arXiv preprint arXiv:1202.3778, 2012.

[16] Deerwester S. C., Dumais S. T., Landauer T. K., Furnas G. W., and Harshman R. A. "Indexing by latent semantic analysis", JASIS, vol. 41, pp. 391-407, 1990.

[17] Hofmann T. "Probabilistic latent semantic indexing", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 50-57, 1999.

[18] Heinrich G. "Parameter estimation for text analysis". Web: http://www. arbylon. net/publications/text-est. pdf, 2005.

[19] Vapnik V. "The nature of statistical learning theory". springer, 2000.