# A Novel Content Enriching Model for Microblog Using News Corpus

**Yunlun Yang**[1], **Zhihong Deng**[2*], **Hongliang Yu**[3]
Key Laboratory of Machine Perception (Ministry of Education),
School of Electronics Engineering and Computer Science,
Peking University, Beijing 100871, China
[1]incomparable-lun@pku.edu.cn
[2]zhdeng@cis.pku.edu.cn
[3]yuhongliang324@gmail.com

## Abstract

In this paper, we propose a novel model for enriching the content of microblogs by exploiting external knowledge, thus improving the data sparseness problem in short text classification. We assume that microblogs share the same topics with external knowledge. We first build an optimization model to infer the topics of microblogs by employing the topic-word distribution of the external knowledge. Then the content of microblogs is further enriched by relevant words from external knowledge. Experiments on microblog classification show that our approach is effective and outperforms traditional text classification methods.

## 1 Introduction

During the past decade, the short text representation has been intensively studied. Previous researches (Phan et al., 2008; Guo and Diab, 2012) show that while traditional methods are not so powerful due to the data sparseness problem, some semantic analysis based approaches are proposed and proved effective, and various topic models are among the most frequently used techniques in this area. Meanwhile, external knowledge has been found helpful (Hu et al., 2009) in tackling the data scarcity problem by enriching short texts with informative context. Well-organized knowledge bases such as Wikipedia and WordNet are common tools used in relevant methods.

Nowadays, most of the work on short text focuses on microblog. As a new form of short text, microblog has some unique features like informal spelling and emerging words, and many microblogs are strongly related to up-to-date topics as well. Every day, a great quantity of microblogs more than we can read is pushed to us, and finding what we are interested in becomes rather difficult, so the ability of choosing what kind of microblogs to read is urgently demanded by common user. Such ability can be implemented by effective short text classification.

Treating microblogs as standard texts and directly classifying them cannot achieve the goal of effective classification because of sparseness problem. On the other hand, news on the Internet is of information abundance and many microblogs are news-related. They share up-to-date topics and sometimes quote each other. Thus, external knowledge, such as news, provides rich supplementary information for analysing and mining microblogs.

Motivated by the idea of using topic model and external knowledge mentioned above, we present an LDA-based enriching method using the news corpus, and apply it to the task of microblog classification. The basic assumption in our model is that news articles and microblogs tend to share the same topics. We first infer the topic distribution of each microblog based on the topic-word distribution of news corpus obtained by the LDA estimation. With the above two distributions, we then add a number of words from news as additional information to microblogs by evaluating the relatedness of between each word and microblog, since words not appearing in the microblog may still be highly relevant.

To sum up, our contributions are:

(1) We formulate the topic inference problem for short texts as a convex optimization problem.

(2) We enrich the content of microblogs by inferring the association between microblogs and external words in a probabilistic perspective.

(3) We evaluate our method on the real datasets and experiment results outperform the baseline methods.

*Corresponding author



*Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 218–223,
Baltimore, Maryland, USA, June 23-25 2014. 

## 2 Related Work

Based on the idea of exploiting external knowledge, many methods are proposed to improve the representation of short texts for classification and clustering. Among them, some directly utilize the structure information of organized knowledge base or search engine. Banerjee et al. (2007) use the title and the description of news article as two separate query strings to select related concepts as additional feature. Hu et al. (2009) present a framework to improve the performance of short text clustering by mining informative context with the integration of Wikipedia and WordNet.

However, to better leverage external resource, some other methods introduce topic models. Phan et al. (2008) present a framework including an approach for short text topic inference and adds abstract words as extra feature. Guo and Diab (2012) modify classic topic models and proposes a matrix-factorization based model for sentence similarity calculation tasks.

Those methods without topic model usually rely greatly on the performance of search system or the completeness of knowledge base, and lack in-depth analysis for external resources. Compared with our method, the topic model based methods mentioned above remain in finding latent space representation of short text and ignore that relevant words from external knowledge are informative as well.

## 3 Our Model

We formulate the problem as follows. Let $EK = \{d_1^e, \ldots, d_{M^e}^e\}$ denote external knowledge consisting of $M^e$ documents. $V^e = \{w_1^e, \ldots, w_{N^e}^e\}$ represents its vocabulary. Let $MB = \{d_1^m, \ldots, d_{M^m}^m\}$ denote microblog set and its vocabulary is $V^m = \{w_1^m, \ldots, w_{N^m}^m\}$. Our task is to enrich each microblog with additional information so as to improve microblog's representation.

The model we proposed mainly consists of three steps:

(a) Topic inference for external knowledge by running LDA estimation.

(b) Topic inference for microblogs by employing the word distributions of topics obtained from step (a).

(c) Select relevant words from external knowledge to enrich the content of microblogs.

### 3.1 Topic Inference for External Knowledge

We do topic analysis for $EK$ using LDA estimation (Blei et al., 2003) in this section and we choose LDA as the topic analysis model because of its broadly proved effectivity and ease of understanding.

In LDA, each document has a distribution over all topics $P(z_k|d_j)$, and each topic has a distribution over all words $P(w_i|z_k)$, where $z_k$, $d_j$ and $w_i$ represent the topic, document and word respectively. The optimization problem is formulated as maximizing the log likelihood on the corpus:

$$\max \sum_i \sum_j X_{ij} \log \sum_k P(z_k|d_j) P(w_i|z_k) \quad (1)$$

In this formulation, $X_{ij}$ represents the term frequency of word $w_i$ in document $d_j$. $P(z_k|d_j)$ and $P(w_i|z_k)$ are parameters to be inferred, corresponding to the topic distribution of each document and the word distribution of each topic respectively. Estimating parameters for LDA by directly and exactly maximizing the likelihood of the corpus in (1) is intractable, so we use Gibbs Sampling for estimation.

After performing LDA model ($K$ topics) estimation on $EK$, we obtain the topic distributions of document $d_j^e$ ($j = 1, \ldots, M^e$), denoted as $P(z_k^e|d_j^e)$ ($k = 1, \ldots, K$), and the word distribution of topic $z_k^e$ ($k = 1, \ldots, K$), denoted as $P(w_i^e|z_k^e)$ ($i = 1, \ldots, N^e$). Step (b) greatly relies on the word distributions of topics we have obtained here.

### 3.2 Topic Inference for Microblog

In this section, we infer the topic distribution of each microblog. Because of the assumption that microblogs share the same topics with external corpus, the "topic distribution" here refers to a distribution over all topics on $EK$.

Differing from step (a), the method used for topic inference for microblogs is not directly running LDA estimation on microblog collection but following the topics from external knowledge to ensure topic consistence. We employ the word distributions of topics obtained from step (a), i.e. $P(w_i^e|z_k^e)$, and formulate the optimization problem in a similar form to Formula (1) as follows:

$$\max_{P(z_k^e|d_j^m)} \sum_i \sum_j \underline{X}_{ij} \log \sum_k P(z_k^e|d_j^m) P(w_i^e|z_k^e), \tag{2}$$

where $\underline{X}_{ij}$ represents the term frequency of word $w_i^e$ in microblog $d_j^m$, and $P(z_k^e|d_j^m)$ denote the distribution of microblog $d_j^m$ over all topics on $EK$. Obviously most $\underline{X}_{ij}$ are zero and we ignore those words that do not appear in $V^e$.

Compared with the original LDA optimization problem (1), the topic inference problem for microblog (2) follows the idea of document generation process, but replaces topics to be estimated with known topics from other corpus. As a result, parameters to be inferred are only the topic distribution of every microblog.

It is noteworthy that since the word distribution of every topic $P(w_i^e|z_k^e)$ is known, Formula (2) can be further solved by separating it into $M^m$ subproblems:

$$\max_{P(z_k^e|d_j^m)} \sum_i \underline{X}_{ij} \log \sum_k P(z_k^e|d_j^m) P(w_i^e|z_k^e) \quad for \quad j = 1, \ldots, M^m \tag{3}$$

These $M^m$ subproblems correspond to the $M^m$ microblogs and can be easily proved convexity. After solving them, we obtain the topic distributions of microblog $d_j^m$ ($j = 1, \ldots, M^m$), denoted as $P(z_k^e|d_j^m)$ ($k = 1, \ldots, K$).

### 3.3 Select Relevant Words for Microblog

To enrich the content of every microblog, we select relevant words from external knowledge in this section.

Based on the results of step (a)&(b), we calculate the word distributions of microblogs as follows:

$$P(w_i^e|d_j^m) = \sum_k P(z_k^e|d_j^m) P(w_i^e|z_k^e), \tag{4}$$

where $P(w_i^e|d_j^m)$ represents the probability that word $w_i^e$ will appear in microblog $d_j^m$. In other words, though some words may not actually appears in a microblog, there is still a probability that it is highly relevant to the microblog. Intuitively, this probability indicates the strength of association between a word and a microblog. The word distribution of every microblog is based on topic analysis and its accuracy relies heavily on the accuracy of topic inference in step (b). In fact, the more words a microblog includes, the more accurate its topic inference will be, and this can be regarded as an explanation of the low efficiency of data sparseness problem.

For microblog $d_j^m$, we sort all words by $P(w_i^e|d_j^m)$ in descending order. Having known the top $L$ relevant words according to the result of sorting, we redefine the "term frequency" of every word after adding these $L$ words to microblog $d_j^m$ as additional content. Supposing these $L$ words are $w_{j1}^e, w_{j2}^e, \ldots, w_{jL}^e$, the revised term frequency of word $w \in \{w_{j1}^e, \ldots, w_{jL}^e\}$ is defined as follows:

$$RTF(w, d_j^m) = \frac{P(w|d_j^m)}{\sum_{p=1}^{L} P(w_{jp}^e|d_j^m)} * L, \tag{5}$$

where $RTF(\cdot)$ is the revised term frequency.

As the Equation (5) shows, the revised term frequency of every word is proportional to probability $P(w_i|d_j^m)$ rather than a constant.

So far, we can add these $L$ **words and their revised term frequency** as additional information to microblog $d_j^m$. The revised term frequency plays the same role as TF in common text representation vector, so we calculate the TFIDF of the added words as:

$$TFIDF(w, d_j^m) = RTF(w, d_j^m) \cdot IDF(w) \tag{6}$$

Note that $IDF(w)$ is changed as arrival of new words for each microblog. The TFIDF vector of a microblog with additional words is called **enhanced vector**.

## 4 Experiment

### 4.1 Experimental Setup

To evaluate our method, we build our own datasets. We crawl 95028 Chinese news reports from Sina News website, segment them, and remove stop words and rare words. After preprocessing, these news documents are used as external knowledge. As for microblog, we crawl a number of microblogs from Sina Weibo, and ask unbiased assessors to manually classify them into 9 categories following the column setting of Sina News.

Sina News: http://news.sina.com.cn/
Sina Weibo: http://www.weibo.com/

After the manual classification, we remove short microblogs (less than 10 words), usernames, links and some special characters, then we segment them and remove rare words as well. Finally, we get 1671 classified microblogs as our microblog dataset. The size of each category is shown in Table 1.

| Category | #Microblog |
|---|---|
| Finance | 229 |
| Stock | 80 |
| Entertainment | 162 |
| Military Affairs | 179 |
| Technologies | 204 |
| Digital Products | 194 |
| Sports | 195 |
| Society | 214 |
| Daily Life | 214 |

Table 1: Microblog number of every category

There are some important details of our implementation. In step (a) of Section 3.1 we estimate LDA model using GibbsLDA++, a C/C++ implementation of LDA using Gibbs Sampling. In step (b) of Section3.2, OPTI toolbox on Matlab is used to help solve the convex problems. In the classification tasks shown below, we use LibSVM as classifier and perform ten-fold cross validation to evaluate the classification accuracy.

### 4.2 Classification Results

| Representation | Average Accuracy |
|---|---|
| TFIDF vector | 0.7552 |
| Boolean vector | 0.7203 |
| **Enhanced vector** | **0.8453** |

Table 2: Classification accuracy with different representations

In this section, we report the average precision of each method as shown in Table 2. The *enhanced vector* is the representation generated by our method. Two baselines are *TFIDF vector* (Jones, 1972) and *boolean vector* (word occurrence) of the original microblog. In the table, our method increases the classification accuracy from 75.52% to 84.53% when considering additional information, which means our method indeed improves the representation of microblogs.

### 4.3 Parameter Tuning

#### 4.3.1 Effect of Added Words

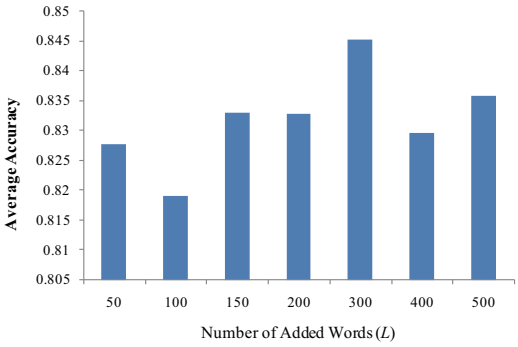


Figure 1: Classification accuracy changes according to topics and added words

The experiment corresponding to Figure 1 is to discover how the classification accuracy changes when we fix the number of topics ($K = 100$) and change the number of added words ($L$) in our method. Result shows that more added words do not mean higher accuracy. By studying some cases, we find out that if we add too many words, the proportion of "noisy words" will increase. We reach the best result when number of added words is 300.

#### 4.3.2 Effect of Topic Number

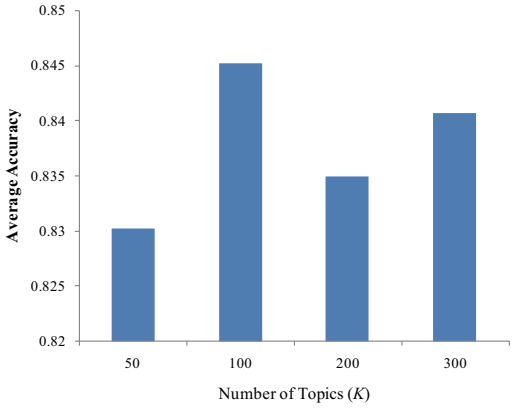


Figure 2: Classification accuracy changing according to the number of topics

The experiment corresponding to Figure 2 is to discover how the classification accuracy changes when we fix the number of added words ($L =$

GibbsLDA++: http://gibbslda.sourceforge.net
OPTI Toolbox: http://www.i2c2.aut.ac.nz/Wiki/OPTI/
SVM.NET: http://www.matthewajohnson.org/software/svm.html

| Microblog (Translated) | Top Relevant Words (Translated) |
|---|---|
| Kim Jong Un held an emergency meeting this morning, and commanded the missile units to prepare for attacking U.S. military bases at any time. | South Korea, America, North Korea, work, safety, claim, military, exercise, united, report |
| Shenzhou Nine will carry three astronauts, including the first Chinese female astronaut, and launch in a proper time during the middle of June. | day, satellite, launch, research, technology, system, mission, aerospace, success, Chang'e Two |

Table 3: Case study (Translated from Chinese)

300) and change the number of topics ($K$) in our method. As we can see, the accuracy does not grow monotonously as the number of topics increases. Blindly enlarging the topic number will not improve the accuracy. The best result is reached when topic number is 100, and similar experiments adding different number of words show the same condition of reaching the best result.

#### 4.3.3 Effect of Revised Term Frequency

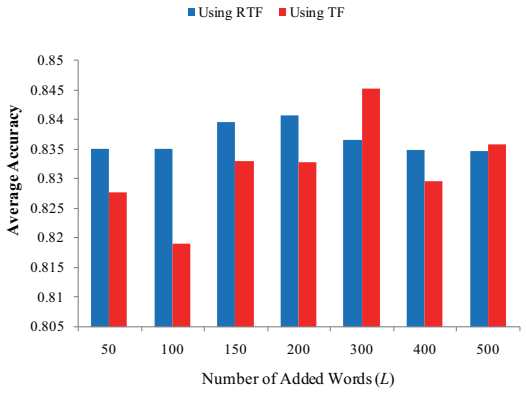


Figure 3: Classification accuracy changing according to the redefinition of term frequency

The experiment corrsponding to Figure 3 is to discover whether our redefining "term frequency" as revised term frequency in step (c) of Section 3.3 will affect the classification accuracy and how. The results should be analysed in two aspects. On one hand, without redefinition, the accuracy remains in a stable high level and tends to decrease as we add more words. One reason for the decreasing is that "noisy words" have a increasing negative impact on the accuracy as the proportion of "noisy words" grows with the number of added words. On the other hand, the best result is reached when we use the revise term frequency. This suggests that our redefinition for term frequency shows better improvement for microblog representation under certain conditions, but is not optimal under all situations.

### 4.4 Case Study

In Table 3, we select several cases consisting of microblogs and their top relevant words .

In the first case, we successfully find the country name according to its leader's name and limited information in the sentence. Other related countries and events are also selected by our model as they often appear together in news. In the other case, relevant words are among the most frequently used words in news and have close semantic relations with the microblogs in certain aspects.

As we can see, based on topic analysis, our model shows strong ability of mining relevant words. Other cases show that the model can be further improved by removing the noisy and meaningless ones among added words.

## 5 Conclusion and Future Work

We propose an effective content enriching method for microblog, to enhance classification accuracy. News corpus is exploited as external knowledge. As for techniques, our method uses LDA as its topic analysis model and formulates topic inference for new data as convex optimization problems. Compared with traditional representation, enriched microblog shows great improvement in classification tasks.

As we do not control the quality of added words, our future work starts from building a filter to select better additional information. And to make the most of external knowledge, better ways to build topic space should be considered.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 61170091).

## References


Banerjee, S., Ramanathan, K., and Gupta, A. 2007, July. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 787-788). ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation. In *Journal of machine Learning research*, 3, 993-1022.

Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007. Measuring semantic similarity between words using web search engines. *www, 7,* 757-766.

Boyd, S. P., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Gabrilovich, E., and Markovitch, S. 2007, January. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI* (Vol. 7, pp. 1606-1611).

Guo, W., and Diab, M. 2012, July. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: Long Papers-Volume 1 (pp. 864-872).

Guo, W., and Diab, M. 2012, July. Learning the latent semantics of a concept from its definition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: Short Papers-Volume 2 (pp. 140-144).

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

Jones, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. In *Journal of documentation*, 28(1), 11-21

Phan, X. H., Nguyen, L. M., and Horiguchi, S. 2008, April. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *Proceedings of the 17th international conference on World Wide Web* (pp. 91-100). ACM.

Sahami, M., and Heilman, T. D. 2006, May. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web* (pp. 377-386). ACM.

Zubiaga, A., and Ji, H. 2013, May. Harnessing web page directories for large-scale classification of tweets. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 225-226). International World Wide Web Conferences Steering Committee.