# Robust Word-Network Topic Model for Short Texts

Fei Wang[†], Rui Liu[†], Yuan Zuo[†*], Junjie Wu[‡]
[†]School of Computer Science and Engineering, Beihang University, Beijing, China
[‡]School of Economics and Management, Beihang University, Beijing, China
[*]*corresponding author:skywatcher.buaa@gmail.com*

***Abstract*—With the rapid development of online social media, the short text has become the prevalent format for information of Internet. Due to the severe data sparsity issue, accurately discovering knowledge behind these short texts remains a critical challenge. Since regular topic models, such as the Latent Dirichlet Allocation (LDA), can not perform well on short texts, many efforts have been put on building different types of probabilistic topic models for short texts. Inducing topics from dense word-word space instead of sparse document-word space becomes an emerging solution for avoiding data sparsity issue, and the representative one is the Word Network Topic Model (WNTM). However, the word-word space building procedure of WNTM often imports much irrelevant information. In light of this, we propose the Robust WNTM (RWNTM), which can filter out unrelated information during the sampling. The experimental results demonstrate that our method can learn more coherent topics and is more accurate in text classification, as compared with WNTM and other state-of-the-arts.**

## I. Introduction

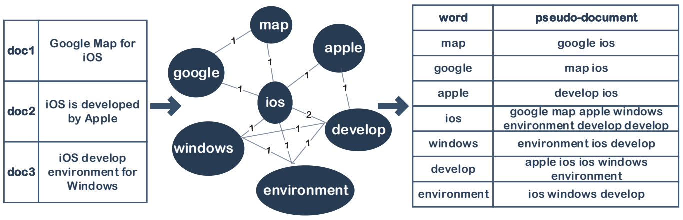
Short texts have been becoming the dominating content of Internet, since the rapid growth of online social media such as Twitter and Facebook. For instance, around 250 million active users in Twitter can generate nearly 500 million tweets everyday. This huge volume of short texts contains sophisticated information, which reflects the real world. Hence, accurately uncover knowledge under these short texts has been recognized as a challenging and promising research task.

Probabilistic topic models have been widely used to automatically extract thematic information from a large archive of documents [1]. Standard topic models [2], [3] assumes that a document is generated from a mixture of topics while a topic is a probabilities distributions over words. In essence, topic models take advantage of document-level word co-occurrence information [4] to form topics. Since short texts contain only a few words in each document, word co-occurrence information is too few for standard topic models to learn coherent topics. Taking LDA as an example, as one of the most typical probabilistic topic models, LDA has achieved great success in modeling regular texts like news articles, research papers and blogs. However, the results are mixed when LDA is applied directly to short texts like tweets, instant messages and forum messages. In order to tackle the incompetence of standard topic models in modeling short texts, many research efforts have been devoted.

One straightforward strategy, widely adopted for short texts in social media, is to utilize auxiliary contextual information to boost the training of topics. For instance, tweets contain not only textual content but also contextual information such as authorship, hashtag, time, location and URL. The auxiliary information can be used to reorganize short texts. Hong and Davison [5] also report better topic model can be trained on aggregated tweets. Mehrotra et al. [6] compare several ways of tweet aggregation using different auxiliary information, and find the one with hashtag yields the best performance. Besides of direct aggregation, some variants of the basic topic model have also been proposed. For example, Tang et al. [7] propose a multi-context topic model, which regularize topics of short texts according to their auxiliary context. Jin et al [8] use the web pages pointed by URLs in tweets as auxiliary long documents to learn better topics in tweets.

In practice, however, the auxiliary information is not always available. Therefore, many research efforts have been put on designing generalized topic models for short texts. For example, Zhao et al [9] assume each tweet only contains one topic, and apply Mixture of unigram (MU) to extract topic from tweets. This way of modeling indeed adds an extreme sparse constraint over document topic distribution. However, given a tweet, it might contain two or three topics, which violates the constraint of MU. Yan et al [10] propose the Biterm Topic Model (BTM), which assumes any two words within a short text come from one topic. They show BTM can learn more coherent topic than MU on short texts. However, BTM is actually a special form of MU, which indicates BTM does not overcome the above issue of MU. Zuo et al. [11] propose the Word Network Topic Model (WNTM) to learn topics from word co-occurrence networks, which produces more coherent topics than BTM.

WNTM converts document-word space to word-word space, for avoiding of document-level word sparsity. Specifically, WNTM utilizes document-level word co-occurrence to build a word co-occurrence network, then generates pseudo-documents from it, finally extracts topics from those pseudo-documents. Although WNTM performs well on topic modeling of short texts, it still has its limitations. Specifically, when building the word network, WNTM often brings in much unrelated word co-occurrence information, which hurts the coherence of its learned topics. Motivated by above consideration, we propose the Robust Word Network Topic Model (RWNTM), which filters out unexpected word co-occurrences information and improves the topic coherence of WNTM.



|  |  |
|---|---|
| doc1 | Google Map for iOS |
| doc2 | iOS is developed by Apple |
| doc3 | iOS develop environment for Windows |

| word | pseudo-document |
|---|---|
| map | google ios |
| google | map ios |
| apple | develop ios |
| ios | google map apple windows environment develop develop |
| windows | environment ios develop |
| develop | apple ios ios windows environment |
| environment | ios windows develop |

Fig. 1. An illustration of the input for RWNTM.

Extensive experiments on two real-world data sets with classic as well state-of-the-art baselines demonstrate the high quality of topics learned by RWNTM.

The rest of this paper is organized as follows: in Section II, we give a brief review of WNTM. Section III introduces the Robust WNTM. Experiment results are presented in Section IV. At last, we give the conclusion.

## II. Word Network Topic Model

In this section, we give some details of the Word Network Topic Model (WNTM) [11], for a better understanding of the Robust WNTM (RWNTM).

When texts are short, document-word space is sparse, while word-word space is still rather dense. WNTM's main idea is converting sparse document-word space to dense word-word space and then extracting topics from word-word space. In WNTM, a word co-occurrence network is used to represent a word-word space. In a word co-occurrence network, nodes are words occurring in the corpus and an edge between two words indicates that the connected two words have co-occurred in the same sliding window at least once.

In order to convert the given document collection into a word co-occurrence network, WNTM applied a sliding window to scan each document. As the window scanning word by word through the document, any two distinct words appear in the same window would be regarded as co-occurred with each other. Times that two words co-occurred are accumulated and defined as the weight of the corresponding edge between them. *Note that Latent word groups in word co-occurrence network are taken as topics in LDA.*

To find these latent word groups from word co-occurrence network, WNTM represents the word co-occurrence network in the form of adjacent lists. As showed in Fig. 1, each word's adjacent list constitutes the content of the corresponding pseudo-document. Compared with the original short texts, those constructed pseudo-documents have a longer length and enriched word co-occurrence information. Which benefits the training of topic models.

For the learning of latent word groups, WNTM assumes a generative process very similar to LDA. It first supposes that there is a fixed set of latent word groups in the word co-occurrence network, and each latent word group $z$ is associated with a multinomial distribution over the vocabulary $\Phi_z$ , which is drawn from a Dirichlet prior $Dir(\beta)$. The generative process of the whole network can be interpreted as follows

1) For each latent word group $z$, draw $\Phi_z \sim Dir(\beta)$
2) For each pseudo-document $l_i$ (or adjacent list of word $w_i$), draw $\Theta_i \sim Dir(\alpha)$
3) For each word $w_j \in l_i$:
    a) Draw a latent word group $z_j \sim \Theta_i$
    b) Draw the adjacent word $w_j \sim \Phi_{z_j}$

When topic proportions of word $w_i$'s adjacent word-list $\Theta_i$ is obtained, topic proportions of original document $d$ can be inferred with following equations

$$p(z|d) = \sum_{w_i} p(z|w_i)p(w_i|d) \tag{1}$$

$$p(w_i|d) = \frac{n_d(w_i)}{Len(d)} \tag{2}$$

where $p(z|w_i)$ equals to $\Theta_{i,z}$, $n_d(w_i)$ is the word frequency of $w_i$ in document $d$ and $Len(d)$ is the length of $d$.

*Remark.* The word-word space that WNTM used enriches the word co-occurrence information. In the meantime, it also brings in some unexpected word co-occurrences, since semantically unrelated words might also co-occur in the same short text. However, WNTM uses the whole adjacent word list of a word without filtering less related words to form its corresponding pseudo-document. Besides, the generative process of WNTM does not take this situation into consideration. Therefore, noises bring into the pseudo-document will limit the accuracy of learned topic distribution of the pseudo-document, which further hurts topic coherence of learned topics.

## III. Robust Word Network Topic Model

As discussed above, semantically unrelated words can be connected in the word co-occurrence network. These connections prevent WNTM from learning more coherent topics. In order to alleviate this situation, one might apply some filtering procedure to remove less related words from one word's adjacent list. However, this method faces two practical issues.

The first issue is how to measure the semantic closeness of two words. Some mostly used measurement such as Pointwise Mutual Information (PMI) does not function well on short texts, due to sparse word co-occurrence. One might compute PMI on auxiliary corpus instead of training short texts. However, for domain specific short texts, the appropriate auxiliary corpus is hard to obtain. The second issue is how to set the threshold of the filtering procedure. Given the PMI score of a pair of words, we need a threshold to determine whether the two words are related or not. However, setting the threshold appropriately is nontrivial.

Different from applying filtering procedure to word's adjacent list, we propose the Robust Word Network Topic Model (RWNTM), which can filter out less related words from topics (or latent word groups) automatically.

WNTM assumes a word's adjacent list is generated by its latent word groups, which are multinomial distributions over the vocabulary. Therefore, unrelated words bring noise into these distributions. We add a so-called specific distribution

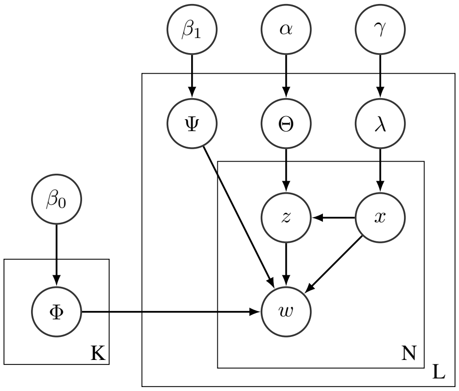


Fig. 2. Plate notation of RWNTM.

for each word's adjacent list, so related words tended to be generated via latent word groups. Since unrelated words themselves do not form semantic coherent topics, they tend to be generated by specific distribution introduced into WNTM. This way of modeling automatically filters out unrelated word connections and the remain related words can form more coherent and more accurate topics.

### A. Generative Process of Robust WNTM

Suppose there are $K$ multinomial distributions $\{\Phi_z\}_{z=0}^{K}$ from a symmetric Dirichlet prior with parameter $\beta_0$, each corresponds to a latent word group (or topic). In additional to latent word groups, there are $L$ multinomial distributions $\{\Psi_l\}_{l=0}^{L}$ from a symmetric Dirichlet prior with parameter $\beta_1$, each corresponds to adjacent list's specific distribution. We introduce an indicator $x$ to decide whether a word is generated from topics or specific distribution. Specifically, when $x = 0$, a word is generated from topics $\Phi$, otherwise, a word is generated from specific distribution $\Psi$. $x$ is drawn from a Bernoulli distribution parameterized by $\lambda$, which in turn is drawn from a symmetric $Beta(\gamma)$. The plate notation is illustrated in Fig. 2. We give the generative process of RWNTM below:

1) For each latent word group $z$:
   a) Draw $\Phi_z \sim Dir(\beta_0)$
2) For each adjacent word-list $l_i$ of the word $w_i$:
   a) Draw $\Theta_i \sim Dir(\alpha)$
   b) Draw $\lambda_i \sim Beta(\gamma)$
   c) Draw $\Psi_i \sim Dir(\beta_1)$
3) For each word $w_j \in l_i$:
   a) Draw $x_j \sim \lambda_i$
   b) If $x_j = 0$ draw $z_j \sim \Theta_i$ and draw $w_j \sim \Phi_{z_j}$
      If $x_j = 1$ draw $w_j \sim \Psi_i$

*Remark.* The introduction of specific distribution $\Psi$ is the key to filtering out irrelevant words from a given adjacent list, which improves the coherence of learned topics $\Phi$.

### B. Inference

Exact posterior inference is intractable in our model. We turn to a collapsed Gibbs sampling algorithm [12] for approximate posterior inference, which is simple to derive, comparable in speed to other estimators, and can approximate a global maximum.

The conditional probability of the $j$th word $w_j$ given the $i$th adjacent word-list $l_i$ can be written as:

$$\begin{aligned} p(w_j|l_i) = p(x_j = 0|l_i) \sum_{k=1}^{K} p(w_j|z_j = k) p(z_j = k|l_i) \\ + p(x_j = 1|l_i) p'(w_j|l_i) \end{aligned} \tag{3}$$

where $p(z_j|l_i)$ is the topic distribution of adjacent word-list $l$. $p(w_j|z_j)$ is the word distribution of topic $z_j$. $p'(w_j|l_i)$ is the special word distribution of adjacent word-list $l_i$. It is relatively straightforward to derive Gibbs sampling equations that allow joint sampling of the $z_j$ and $x_j$ latent variables for each word $w_j$, for $x_j = 0$:

$$\begin{aligned} p(x_j = 0, z_j = k|\mathbf{w}, \mathbf{x}_{-j}, \mathbf{z}_{-j}, \alpha, \beta_0, \gamma) \propto \\ \frac{N_{0,-j}^{l_i} + \gamma}{N_{0,-j}^{l_i} + N_{1,-j}^{l_i} + 2\gamma} \times \frac{N_{k,-j}^{l_i} + \alpha}{\sum_{k'} N_{k',-j}^{l_i} + K\alpha} \\ \times \frac{N_{w,-j}^{k} + \beta_0}{\sum_{w'} N_{w',-j}^{k} + V\beta_0} \end{aligned} \tag{4}$$

for $x_j = 1$:

$$\begin{aligned} p(x_j = 1|\mathbf{w}, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \alpha, \beta_1, \gamma) \propto \frac{N_{1,-j}^{l_i} + \gamma}{N_{0,-j}^{l_i} + N_{1,-j}^{l_i} + 2\gamma} \\ \times \frac{N_{w,-j}^{l_i} + \beta_1}{\sum_{w'} N_{w',-j}^{l_i} + V\beta_1} \end{aligned} \tag{5}$$

where the subscript $-j$ indicates that the count for $j$th word is removed, $N_0^{l_i}$ and $N_1^{l_i}$ are the number of words in $ith$ adjacent list $l_i$ that assigned to the latent topics and special words respectively. $N_k^{l_i}$ is the number of words in adjacent list $l_i$ that assigned to topic $k$. $N_w^k$ and $N_w^{l_i}$ are the number of times word $w$ is assigned to topic $k$ and to the special words distribution of adjacent word-list $l_i$ respectively. $K$ is the number of all latent topics and $V$ is the number of unique words in the corpus.

## IV. Experiments

In this section, we present extensive experimental results on two real-world data sets to evaluate our model according to topic coherence scores and document classification results.

### A. Experimental Setup

*1) Date sets:* Two real-world short texts with statistic listed in Table I.

**News.** This data set[1] contains 29,200 English news articles extracted from RSS feeds of three popular newspaper websites (nyt.com, usatoday.com, reuters.com). Categories include

[1] http://acube.di.unipi.it/tmn-dataset/

TABLE I
STATISTICS OF DATA SETS.

| Data set | # Documents | Vocabulary size | Avg. document length |
|---|---|---|---|
| News | 29,200 | 11,007 | 12.4 |
| Tweets | 182,671 | 21,480 | 8.5 |

sport, business, U.S., health, sci&tech, world and entertainment. We retain news descriptions since they are typical short texts.

**Tweets.** A large set of tweets are collected and labeled by Zubiaga et al. [13]. They crawl tweets that contain URL and label them with the categories of web pages pointed by the URLs. The categories of web pages are defined by the Open Directory Project (ODP). This dataset contains ten different categories and totally around 360k labeled tweets. We select nine topic-related categories and sample 182,671 tweets in total under those categories.

*2) Baseline Methods:* Four baseline methods are included for a thorough comparative study.

**Latent Dirichlet Allocation (LDA) [2].** Being one of the most classical topic model. We use jGibbLDA[2] package with collapsed Gibbs sampling.

**Mixture of Unigram (MU).** The most important feature of MU [14], [15] is that it assumes each document is generated by only one topic, which may be feasible for certain collections of short texts.

**Biterm Topic Model (BTM) [10].** BTM is a topic model specific for short text. It trains topics on co-occurred word pairs. The code we used was downloaded from github[3] which was upload by the author of BTM.

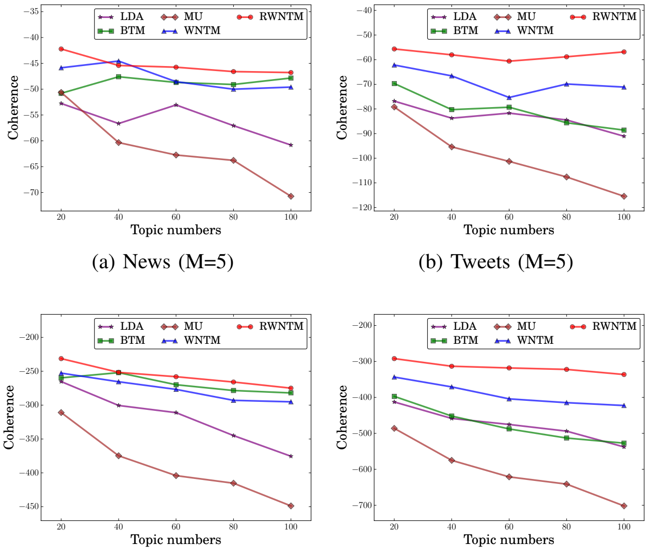
**Word Network Topic Model (WNTM) [11].** Since our model is rooted from WNTM, comparison with WNTM can directly show the effectiveness of our improvements of WNTM.

*3) Evaluation Measures:* We apply UMass topic coherence [16] to evaluate the topic quality and document classification to evaluate latent semantic representations learned by our model.

**Topic Coherence**. Evaluation of topic models is still an open problem. The commonly used metric named *perplexity* has been proved less correlated to human interpretability. Many methods thus turn to use *topic coherence* to evaluate topics, which is proved more correlated to human evaluations.

UMass Topic coherence [16] is a measurement that evaluates quality of topics. Given a topic $z$ and $T^{(z)} = (w_1^{(z)}, ..., w_M^{(z)})$ is a list of the $M$ most probable words of topic $z$, ordered by $P(w|z)$. The coherence score of topic $z$ is defined as:

$$C(z; T^{(z)}) = \sum_{m=2}^{M} \sum_{n=1}^{m} log \frac{D(w_m^{(z)}, w_n^{(z)}) + \epsilon}{D(w_n^{(z)})} \tag{6}$$



(a) News (M=5)

(b) Tweets (M=5)

(c) News (M=10)

(d) Tweets (M=10)

Fig. 3. Coherence scores on News and Tweets.

where $D(w)$ is the document frequency of word $w$ and $D(w, w')$ is the number of documents words $w$ and $w'$ co-occurred. $\epsilon = 10^{-12}$ was set to avoid taking the log of zero for words that never co-occur and to smooth the score for completely unrelated words [17]. And the final coherence score is the average coherence score of all topics. The coherence score is based on the idea that words belonging to a single concept will tend to co-occur within the same documents. Higher topic coherence often indicates better topic quality.

**Classification measures.** Topic models are often evaluated on external tasks such as text classification. Therefore, we also conduct short-text classification experiments to compare the latent semantic representations learned by our methods and baselines. Macro-averaged f-measure are used in classifications.

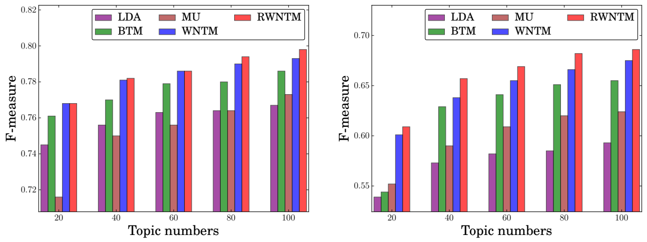
For our model, we set $\alpha = 0.1$, $\beta_0 = 0.01$, $\beta_1 = 0.0001$, $\gamma = 0.3$. And we set $\alpha = 0.1$, $\beta = 0.01$ for WNTM, which are suggested by [11]. For LDA, we manually set $\alpha = 0.1$ and $\beta = 0.01$, since LDA with weak priors performs better in short texts. For MU, we also set $\alpha = 0.1$ and $\beta = 0.01$. For BTM, we set $\alpha = 0.5$, $\beta = 0.005$, which are default settings in [10].

## B. Experimental Results

*1) Topic Coherence Results:* The UMass topic coherence results of our method and all baselines on News and Tweets are presented in Fig. 3. From the results, we can find MU has the lowest coherence scores. LDA performs better than MU. The worse topic coherence of MU and LDA illustrates the uncompetitive of classical topic models on short texts. BTM outperforms LDA on News, while achieves similar coherence scores compared to LDA on Tweets. Similarly, BTM slightly outperforms WNTM on News, while WNTM achieves better coherence scores than BTM on Tweets. WNTM achieves the

[2] http://jgibblda.sourceforge.net/

[3] https://github.com/xiaohuiyan/BTM



(a) News

(b) Tweets

Fig. 4. F-measure results on News and Tweets.

best overall topic coherence compared to other three baselines. This indicates the advantage of word-word spaces to learn more coherent topic on short texts. RWNTM achieves the best coherence scores in almost every cases, except the result on News, when the number of topic equals to 40. Which in consistence with our consideration that noise connections limits WNTM's ability of learning more coherent topics.

*2) Classification Results:* Taking topic model as a method of dimensionality reduction, we can reduce a document into a fixed set of topics. Document classification results indicate the discriminative ability of topic representations of documents learned by models.

Since RWNTM, WNTM and BTM do not model topics of documents directly, they have to resort to certain post inference strategies to obtain indirect representations. Empirically, indirect representation gives better classification result than direct representations. Thus, for a fair comparison, we obtain indirect representation of documents for LDA and MU. Specifically, we apply the follow post inference strategy to represent the document $d$. $p(z|d) = \sum_w p(z|w)p(w|d)$, where $p(z|w)$ is estimated by $\frac{p(z)p(w|z)}{\sum_z p(z)p(w|z)}$ and $p(w|d)$ is estimated using the relative frequency of $w$ in $d$.

For all methods, we perform five-fold cross validation. In each fold, we randomly split data set into training and test subsets with the ratio 4:1 and classify them using LIBLINEAR[4]. All results are illustrated in Fig. 4.

From the results we can find RWNTM consistently outperforms other methods according to f-measure on both News and Tweets, with varying number of topics. The outperformance of RWNTM as compared to WNTM shows that the introduction of specific distributions boosts the quality of topics, which further guarantees the topic representation of a document. WNTM outperforms other baseline methods, which indicates word-word space has advantages in learning semantic representations of short texts. BTM outperforms LDA and MU. BTM directly models word pairs, which avoid the learning of topic distributions of short texts. Therefore, BTM is less influenced by sparsity issue of short texts. Interestingly, MU outperforms LDA on Tweets, while is less competitive on News. The reason might be each tweet often contains one topic, which perfectly match the assumption of MU. Overall, this result illustrates the superiority of WNTM to other baseline methods, and the effectiveness of RWNTM as compared to WNTM.

## V. Conclusion

In this paper, we propose a Robust Word Network Topic Model (RWNTM) for short texts. By adding specific distributions to each word's adjacent list, RWNTM can automatically filter out less related connections of that word. In this way, quality of learned topics can be improved. Extensive experiments on real-world data sets demonstrate the superiority of RWNTM to some state-of-the-art methods.

## References

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

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

[4] J. L. Boyd-Graber and D. M. Blei, "Syntactic topic models," in *Advances in neural information processing systems*, 2009, pp. 185–192.

[5] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 80–88.

[6] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 889–892.

[7] J. Tang, M. Zhang, and Q. Mei, "One theme in all views: Modeling consensus topics in multiple contexts," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 5–13.

[8] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 775–784.

[9] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[10] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 1445–1456.

[11] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, 2016.

[12] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[13] A. Zubiaga and H. Ji, "Harnessing web page directories for large-scale classification of tweets," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 225–226.

[14] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[15] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 233–242.

[16] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[17] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12, 2012, pp. 952–961.

[4] http://www.csie.ntu.edu.tw/cjlin/liblinear/