Topic Modeling for Short Texts with Co-occurrence Frequency-based Expansion

Gabriel Pedrosa*, Marcelo Pita*‡, Paulo Bicalho*, Anisio Lacerda† and Gisele L. Pappa*
*Universidade Federal de Minas Gerais, Belo Horizonte, Brasil
†Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Brasil
‡Serviço Federal de Processamento de Dados, Belo Horizonte, Brasil
Emails: *{gabrielmip, glpappa, marcelo.pita, p.bicalho}@dcc.ufmg.br, †anisio@decom.cefetmg.br

***Abstract*—Short texts are everywhere on the Web, including messages in social media, status messages, etc, and extracting semantically meaningful topics from these collections is an important and difficult task. Topic modeling methods, such as Latent Dirichlet Allocation, were designed for this purpose. However, discovering high quality topics in short text collections is a challenging task. This is because most topic modeling methods rely on information coming from the word co-occurrence distribution in the collection to extract topics. As in short text this information is scarce, topic modeling methods have difficulties in this scenario, and different strategies to tackle this problem have been proposed in the literature. In this direction, this paper introduces a method for topic modeling of short texts that creates pseudo-documents representations from the original documents. The method is simple, effective, and considers word co-occurrence to expand documents, which can be given as input to any topic modeling algorithm. Experiments were run in four datasets and compared against state-of-the-art methods for extracting topics from short text. Results of coherence, NPMI and clustering metrics showed to be statistically significantly better than the baselines in the majority of cases.**

I. Introduction

Topic modeling methods are designed to find semantically meaningful topics from a collection of documents. Topics are usually treated as hidden variables that explain observable ones. Observable variables, in the case of text collections, are the documents and the words that compose them. A widely used and consolidated topic modeling technique is the Latent Dirichlet Allocation (LDA) [1]. LDA finds topics that are hidden in documents by exploiting the co-occurrence of their words. For this reason, LDA usually does not perform well in a few scenarios, including: (i) collections that have few documents; (ii) collections that have too many topics; or (iii) collections in which documents are too short [2]. This paper deals with the problems related to the latter scenario.

Short texts are everywhere on the Web, and topic modeling finds a lot of applications in this context. Nevertheless, discovering high quality topics in short text collections is challenging. The difficulty is due to the high sparsity of the *docs*×*words* matrix, i.e., documents contain only a few words from all available in the collection vocabulary. There are two known approaches that address the problem of topics extraction from short text, namely: (i) methods that propose new probabilistic topic models or modify the traditional Latent Dirichlet Allocation (LDA) algorithm [3], [4], [5]; and (ii) methods that create larger pseudo-documents from short text documents, decreasing the *docs*×*words* matrix sparsity by increasing the number of different words in the documents. These pseudo-document are then given as input to current topic modeling methods [6], [7]. The latter approach has the advantage of being simpler and method-independent, since it only transforms the input data, and is the foundation of the proposed method.

In general, current methods that generate larger pseudo-documents use information about the application's context, and cannot be easily generalized. In [7], for example, the authors propose different tweet pooling schemes and found out that grouping tweets by hashtags is an effective approach to generate good larger pseudo-documents. However, in scenarios where there is not an available common element to merge the documents (e.g. hashtags), this method cannot be directly applied.

This paper proposes a method for document expansion called Co-occurrence Frequency Expansion (CoFE), which is context-independent and allows the user to specify the maximum desired size of the generated pseudo-documents. CoFE exploits the co-occurrence frequency (co-frequency) of terms in the collection in a way that words with high co-frequency have also a high probability of belonging to the same topic. These words are then used to expand the documents, increasing the word co-frequency in the *docs*×*words* matrix.

We compare the results of the proposed strategy with LDA and two other state of the art methods designed for short text: (i) Biterm Topic Modeling (BTM) [8] and (ii) Latent Feature-LDA (LFLDA) [4]. We evaluate CoFE and the baselines using two strategies. The first directly measures the quality of the topics found using two standard metrics for topic quality: topic coherence and NPMI. The second evaluates the performance of the algorithms in the document clustering task. The results show that CoFE obtained the best overall results for the topic quality metrics and similar document clustering performance to state-of-the-art methods.

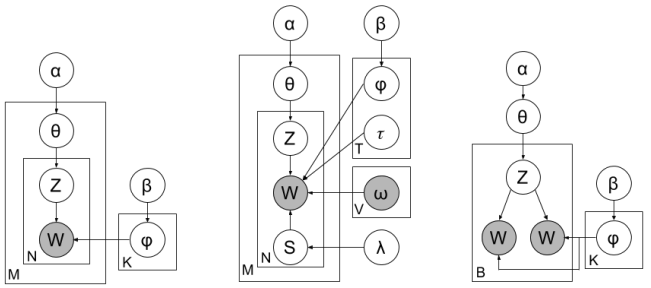
This document is organized as follows. Section II introduces related work on topic modeling for short text. Section III describes CoFE. Section IV introduces the experimental methodology and shows the results obtained. Finally, section V lists our conclusions and future work.

II. Related Work

We first look at approaches for topic modeling in short text that increase the length of the original documents, which mostly focus on Twitter data. For example, Hong et al. [6] proposed two tweet pooling criteria. The first groups tweets by their authors and the second, by vocabulary terms, with the purpose of inferring the topic distribution for authors and tweets. A similar approach was explored by Mehrotra et al. [7], where four tweet pooling schemes were evaluated. In these schemes, tweets were grouped according to the author of the message, the time messages were posted, same hashtags and trending topics. They found out that pooling tweets by hashtags yields the overall best results.

Regarding more general approaches for topic modeling that modified the original LDA method, Zhao et al. [3] proposed Twitter-LDA, which is a modified version of LDA for Twitter with the generative restriction of one topic per tweet. Jin et al [5] proposed Dual LDA (DLDA), which is a method for enhancing short text topic modeling that uses knowledge from an auxiliary dataset of longer texts. Other works have followed similar approaches to DLDA, but ignoring inconsistencies between target and auxiliary data [9].

The current state of the art methods for extracting topics from short text are Latent Feature LDA (LFLDA) [4] and Biterm Topic Modeling (BTM) [8], both generative probabilistic graphical models like LDA. Figure 1 shows the differences between LDA, LFLDA, and BTM models in plate notation. In this notation, labeled boxes indicate replication of variables. M is the number of documents in the collection and N the number of words per document.



(a) LDA (b) LFLDA (c) BTM

Fig. 1. LDA, LFLDA, and BTM graphical models in plate notation.

In the generative perspective of LDA, K latent topics are defined as distributions over the vocabulary words (φ distribution), documents are mixtures of topics (θ distribution) and words (W) are derived, one by one, from topics (according to the Z distribution of topics per word). Only words within documents are observable variables and all priors (α and β) are defined to be Dirichlet distributions. The main objective of the algorithm is to infer the latent variables, such as topics proportions for documents and words distributions per topic.

LFLDA is based on LDA and includes word embedding [10] from a large external corpus, ω, with a vocabulary of size V. Words embeddings are continuous dense vector representations of words in low dimension spaces (relative to V) that are expected to be semantically consistent regarding vector similarity metrics. LFLDA also introduces a latent feature component τ for each of the T topics, which are vector representations of topics learned. For each document d, it draws a multinomial distribution θ_d over all topics. For each ith word w_i in d, it draws a topic indicator z_i and a binary switch s_i. The topic indicator z_i defines from which topic the word w_i is to be generated, and the binary switch s_i determines whether to use the traditional Dirichlet multinomial or the word embeddings.

The generative model of BTM directly models the production of biterms, or word co-occurrence patterns in the same document, and hence addresses the problem of sparsity by grouping biterms. Due to this aggregation, BTM has a single topic distribution for the entire corpus (B biterms), instead of one distribution per document.

III. Co-occurrence Frequency Expansion (CoFE)

The proposed method expands the documents in a collection by appending words considered similar to those already present in the original documents. It is based on the intuition that similar words have higher likelihood of occurring in the same context. In other words, the conditional probability that one word occur in a document, given that a second word was already observed, should be higher if they are similar and lower otherwise. Given a collection of documents D and a maximum document size L, the expansion procedure follows the steps presented in Algorithm 1.

Algorithm 1 Co-Frequency Expansion

Require: D, h, L
1: $G \leftarrow$ Generate words co-frequency graph
2: **for** $d \in D$ **do**
3: **if** $|d| < L$ **then**
4: $G_d \leftarrow$ Extract subgraph
5: $C_d \leftarrow \emptyset$ ▷ Candidate words
6: **for** $e_d = (s, c, w) \in E_d$ **do**
7: $C_{d,c} \leftarrow C_{d,c} + \{w\}$
8: **for** $c \in C_d$ **do**
9: $C_{d,c} \leftarrow sum(c)$
10: **while** $|d| < L$ **do**
11: $h \leftarrow$ SelectionMethod(C_d) ▷ Selected word
12: $d \leftarrow d \cup h$

From the documents D, a word co-frequency graph $G = (N, E)$ is generated (line 1), where N is the set of nodes representing the vocabulary words and E is the set of graph edges. Each node is linked to its h most similar word nodes, where h is a user-defined parameter. To determine how similar any pair of words w_i and w_j are, we used the Jaccard index, defined as:

$$similarity(w_i, w_j) = Jaccard(w_i, w_j) = \frac{|O_i \cap O_j|}{|O_i \cup O_j|} \quad (1)$$

where the sets O_i and O_j contain all documents where words i and j occur, respectively. Word pairs with high co-frequency will have high values of Jaccard index. Graph edges are then weighted by these similarities.

Having the word similarity graph, for each document $d \in D$ with less than L words, the word co-frequency subgraph $G_d = (N_d, E_d)$ is extracted from G, where N_d is the set of nodes representing the document words and new words they are similar to, and E_d the set of graph edges. E_d connects a source word s in the original document to a candidate word c, weighted by the similarity w of s and c. The set of candidate words for expanding document d, C_d, includes all neighbor nodes of the words s (lines 6-7). For each candidate word, we sum up the weights of their in/out degrees (lines 8-9). These final weights are considered by a probabilistic word selection method, which adds the selected word to the original document (lines 10-12), while its size is smaller than the desired maximum size L. The selection is probabilistic to avoid raising the co-occurrence frequency of word pairs excessively, once LDA showed to be very sensible to it.

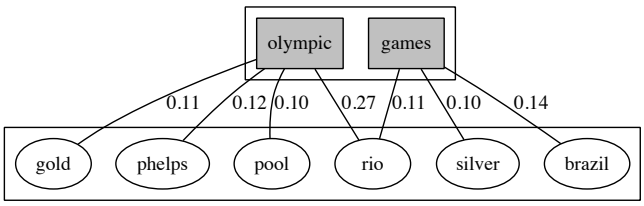


Fig. 2. Example of similarity graph for CoFE ("Olympic Games").

Figure 2 presents an example of a CoFE sub-graph. The original document contains the text "Olympic Games", and for each word we present its neighbor candidate words considering $h = 4$. Assume that the expanded pseudo-document should contain a maximum number of $L = 5$ words after expansion. To determine the probability of a word being selected for expansion, we add up the weights of each edge connected to it, e.g., for *rio* the probability is 0.38 (0.27+0.11). After expansion, the words *rio*, *brazil* and *phelps*, are added in the pseudo-document, which increased its size from 2 to 5, as the original words are kept in the document.

Regarding the time complexity of the method, let N be the number of documents of a dataset, V its vocabulary size and L the expected number of words in an expanded document. In terms of computational complexity, CoFE's cache generation time complexity is of order $O(NV^2)$. The document expansion step has time complexity $O(NLV)$.

IV. Experiments and Results

A. Datasets

We used four short text datasets with labeled documents (required for the clustering evaluation) in our topic modeling experiments:

TABLE I
Characteristics of the datasets.

Dataset	N. of Docs	Vocab. Size	Unique words/doc
20N	1723	964	7.1 (±2.9)
TMN	30376	6314	4.9 (±1.5)
Sanders	3770	1311	5.8 (±2.5)
Snippets	12117	4677	10.3 (±3.1)

1) Tweets Sanders: Tweets related to four different companies: Apple, Google, Microsoft, Twitter.[1]
2) 20 Newsgroups (20N): A collection of documents from 20 newsgroups. We only use the documents with less than 21 words, as done in [4].
3) Tag My News (TMN): A collection of English RSS news items grouped into 7 categories, where only the news titles are considered [11].
4) Web Snippets: A collection of web search snippets, which are summaries of documents presented as results of a query in a search engine [9]. The queries used are related to 8 different domains.

All datasets were preprocessed before the expansion step by making all the text lower-case, removing non-alphabetic characters and stop words. We also removed words shorter than 3 characters and words appearing less than 10 times in 20N and under 5 times in TMN and Sanders.

Table I shows statistics for the datasets after the preprocessing step. For the number of unique words per document, we present the average followed by the standard deviation. Note that the number of unique words per document is low.

B. Parameter Configuration

LDA, LFLDA and BTM share four main parameters: the number of topics (k), the hyper-parameters α and β for the Dirichlet distribution and the number of Gibbs Sampling iterations. The values of α and β for LDA were estimated using Minka's fixed point iteration technique [12], and LDA was run for 2000 iterations. The number of topics assumed values 20, 50 and 100. LFLDA has two extra parameters: the word vector representations and a mixture factor λ, which controls whether to use the Dirichlet or the latent feature component of the method. We use the default value of λ suggested by the authors (0.6), and word vectors learned from Wikipedia (dump 02/06/2015).

We also tested CoFE with different values of L, which corresponds to the target pseudo-document size. We tested four values: 30, 40, 50 and 60 words. The expanded datasets were given to LDA for a previous evaluation of L's influence over LDA performance. The value $L = 60$ showed better overall results for the topic quality evaluation metrics described in Subsection IV-C, being used in further experiments.

C. Quantitative Topic Evaluation

Automatic evaluation of topic modeling methods is not a straightforward process. Here we use two metrics of

[1] Available at http://www.sananalytics.com/lab.

topic quality: the Normalized Pointwise Mutual Information (NPMI)-score [13] and topic coherence [14].

The topic coherence metric evaluates the topic quality by looking at the co-occurrence of the most probable words for the topic in the original dataset. Given a topic t, its 10 most probable words W_{10}, $D(w_i, w_j)$ being the co-document frequency of words w_i and w_j and $D(w_j)$ the document frequency of word w_j, the coherence score for t is:

$$coherence(t; W_{10}) = \sum_{i=2}^{10} \sum_{j=1}^{i-1} log \frac{D(w_i, w_j) + 1}{D(w_j)} \quad (2)$$

The PMI-score [15], in turn, verifies if the semantic relationship between a pair of words suggested by a topic model is also found in an external dataset by evaluating the pointwise mutual information (PMI) of all pairs of its most likely words. The probabilities are evaluated by counting word co-occurrence frequencies in a 10-word sliding window in a large external dataset. Its normalized version was proposed by [13], and removes the score sensitivity to frequency and provides more intuitive score values: when w_i and w_j only occur together, NPMI$(w_i, w_j) = 1$; when they never occur together, NPMI$(w_i, w_j) = -1$. The external dataset used for evaluation consisted of a randomly generated sample of 15M documents in English from the WMT11 news corpus.[2] Given a topic t and its 10 most probable words W_{10}, NPMI-score is defined as:

$$\text{NPMI-Score}(t; W_{10}) = mean\{\text{NPMI}(w_i, w_j), i, j \in 1...10, i \neq j\} \quad (3)$$

$$\text{NPMI}(w_i, w_j) = \left(\ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right) \Big/ - \ln p(w_i, w_j) \quad (4)$$

Table II shows the values of both metrics for the baselines using the original version of the dataset and for LDA using the datasets expanded by CoFE for 20, 50 and 100 topics. Results are compared using a non-parametric Wilcoxon signed-rank test with 0.05 of significance level. Values in bold in the same column (for a dataset) indicate the best methods (one or more). Columns without bold values have no winner method. Note that the values obtained by CoFE are statistically significantly better than the values of all baselines for coherence. This is not a surprise, as the method improves word co-occurrence, which is used by coherence. The second best results are those found by BTM. For the second metric, NPMI, the results are disagree with those of coherence. CoFE was better than the other methods for the Snippets dataset. For the other datasets, there is no method which clearly outperforms the others.

As CoFE is method-independent, we also verified whether the improvements observed for LDA also happen when we give the expanded datasets as input to LFLDA and BTM, as their authors also show they are effective when dealing with larger documents. Table III shows the average values of coherence and NPMI over all datasets when extracting 20 topics, followed by the percentage of improvement over the same method with the original dataset. The expanded datasets considered a maximum document size of 60 words. Looking at the results, we observe that, in these cases, CoFE datasets improve the values of both metrics for all methods, showing that CoFE does not depend on a specific topic modeling algorithm.

[2]Available at http://www.statmt.org/wmt11/training-monolingual.tgz.

TABLE II
RESULTS OF BASELINES RUN WITH THE ORIGINAL DOCUMENTS AND LDA WITH EXPANDED DOCUMENTS.

Topic	20 Topics		50 Topics		100 Topics	
Model	Coherence	NPMI	Coherence	NPMI	Coherence	NPMI
Tweets Sanders						
LDA	-132.1	**-0.087**	-122.6	**-0.099**	-122.8	-0.116
LFLDA	-133.7	**-0.079**	-133.6	**-0.107**	-137.5	-0.143
BTM	-121.9	**-0.085**	-121.5	**-0.097**	-121.8	**-0.109**
LDA-CoFE	**-89.5**	-0.128	**-84.5**	-0.153	**-104.7**	-0.158
20 Newsgroups						
LDA	-111.6	-0.184	-103.4	-0.188	-97.3	**-0.193**
LFLDA	-115.4	-0.179	-113.0	-0.199	-112.716	-0.219
BTM	-106.4	-0.202	-103.1	-0.205	-100.5	-0.208
LDA-CoFE	**-93.9**	-0.194	**-89.3**	-0.199	**-89.4**	-0.200
Tag My News						
LDA	-171.1	-0.062	-161.8	**-0.056**	-156.6	-0.085
LFLDA	-175.5	-0.039	-170.6	**-0.038**	-165.9	**-0.065**
BTM	-169.9	-0.048	-159.5	**-0.049**	-154.2	**-0.069**
LDA-CoFE	**-157.6**	-0.040	**-157.6**	-0.067	**-144.0**	-0.107
Web Snippets						
LDA	-149.6	-0.061	-143.0	-0.102	-133.5	-0.106
LFLDA	-149.5	-0.061	-147.9	-0.094	-144.6	-0.123
BTM	-139.1	**-0.042**	-132.8	-0.082	-124.0	**-0.087**
LDA-CoFE	**-135.0**	**-0.024**	**-126.0**	**-0.054**	**-118.7**	**-0.079**

TABLE III
AVERAGE RESULTS OVER ALL DATASETS FOR LDA, LFLDA AND BTM.

	Original		CoFE	
	Coherence	NPMI	Coherence	NPMI
LDA	-139.5	-0.084	-121.1 (+13.20%)	-0.081 (+3.42%)
LFLDA	-142.7	-0.075	-130.1 (+8.83%)	-0.065 (+13.41%)
BTM	-132.5	-0.083	-122.7 (+7.44%)	-0.077 (+7.24%)

D. Document Clustering Evaluation

This section shows a second type of evaluation of the topics found, namely document clustering. In this scenario, each topic z is considered a cluster and each document d is assigned to the topic with the highest value of conditional probability $P(z \mid d)$.

Let $T_i \in T$ be the set of documents assigned to topic i and let $c_j \in C$ be the set of documents labeled with class j. We consider as classes the datasets document labels: 4 companies for *Sanders*, the 20 newsgroups for *20N*, 7 news categories for TMN and, 8 search domains for Snippets.

To measure the performance of the baselines and CoFE, we used two standard metrics for clustering evaluation:

- *Normalized mutual information* (NMI): NMI is evaluated using the mutual information of T and C, defined by $I(T, C)$. $I(T, C)$ is penalized by the entropy of T and

C, defined as $H(T)$ and $H(C)$, avoiding the bias of a large number of clusters. NMI ranges from 0 to 1, being 1 when the set of cluster labels matches perfectly the document classes. Formally NMI is defined as:

$$\text{NMI}(T, C) = \frac{2I(T, C)}{H(T) + H(C)}$$

- *Adjusted Rand index* (ARI): ARI does a pairwise comparison of the documents in the partitions T and C. Rand index counts the number of times both partitions agree that a document pair should or should not belong to the same cluster, dividing it by the total number of document pairs. Adjusted Rand index is the corrected-for-chance version of Rand index, and ensures an index value to be close to 0 for random labeling and exactly 1 when both partitions always agree. It is defined as:

$$\text{ARI}(T, C) = \frac{\sum_{i,j} \binom{|T_i \cap C_j|}{2} - [\sum_i \binom{|T_i|}{2} \sum_j \binom{|C_j|}{2}]/\binom{M}{2}}{\frac{1}{2}[\sum_i \binom{|T_i|}{2} + \sum_j \binom{|C_j|}{2}] - [\sum_i \binom{|T_i|}{2} \sum_j \binom{|C_j|}{2}]/\binom{M}{2}}$$

Figure 3 shows the results found by the methods for ARI and NMI when we set the number of topics to 20 and 100. Although at first it is unclear which method performs the best in clustering, it is important to point out that CoFE improved LDA's performance by an average of 7.22% for NMI and 70.07% for ARI.

We also compare the number of times each method was statistically significantly the best or present no statistical difference from those with better scores. This comparison is shown in Table IV. For NMI, BTM performs better than LDA-CoFE, while for ARI, our proposed method achieved better results. Overall, LDA-CoFE and BTM perform similarly, being better in 13 and 11 different configurations, respectively.

TABLE IV
NUMBER OF TIMES EACH METHOD WAS STATISTICALLY THE BEST OR PRESENTED NO STATISTICAL DIFFERENCE TO OTHER METHODS.

	Metrics		
Method	NMI	ARI	Total
LDA	4	1	5
LFLDA	1	0	1
BTM	8	3	11
CoFE	4	9	13

E. Topics at a glance

To provide a better insight on how CoFE's document expansion process influences LDA, we compare topics from models trained with the original Tag My News dataset and its expanded version. A good topic should be interpretable and, for this dataset, reflect the news subjects, which include sports, business, health, U.S., science and technology, world and entertainment. Table V shows the topics learned when the number of topics is set to 20. We present the 10 most probable words for each topic. Topics are paired by their cosine similarity using a greedy strategy, which is indicated in column *sim*.

Topics 1-16 present high cosine similarity and each of them can easily be labeled as one of the dataset subjects mentioned above. This suggests that LDA discovered a similar topical structure for both the original and expanded datasets. Topics 17-20, however, present lower similarity, indicating differences in the two models. For the original dataset, topics 17 and 19 do not clearly belong to one specific category. On the other hand, for the expanded dataset, topics 17-19 present higher interpretability and allow easier labeling as health and entertainment. For topic 18, although the topic produced from the original dataset is related to sports, notice that the words produced by CoFE are more discriminative than the one produced from the original dataset. From the results, we conclude that CoFE produce topics with words that are more context-related than the original dataset.

V. CONCLUSION AND FUTURE WORK

This paper introduced CoFE, a simple, effective and efficient method for generating larger pseudo-documents from short-texts. CoFE explores word co-occurrence to expand the text with words similar to the document context. The pseudo-documents produced can be given as input to any method for topic modeling, improving the topics found.

The method was evaluated in four datasets and compared to other state-of-the-art algorithms for topic modeling in short texts. Results show that CoFE produced topics with words that are more context-related than the original dataset. Topic quality, evaluated with coherence and NPMI, also improves significantly when CoFE is given as input to current methods. Regarding document clustering, when evaluated with NMI and ARI, CoFE also improves LDA's performance, being comparable to state-of-the-art methods.

As future work, we intend to further evaluate the performance of the methods as we increase the size of the documents and perform a more complete qualitative evaluation of the topics. Further discussion on which metric is more appropriate to scenarios with short text are also promising directions.

VI. ACKNOWLEDGMENTS

The authors would like to thank CAPES, CNPq and Fapemig, all Brazilian Research Funding agencies, W3C/NIC.br, CEFET-MG/PROPESQ-10314/14 and SERPRO, for their financial support.

REFERENCES

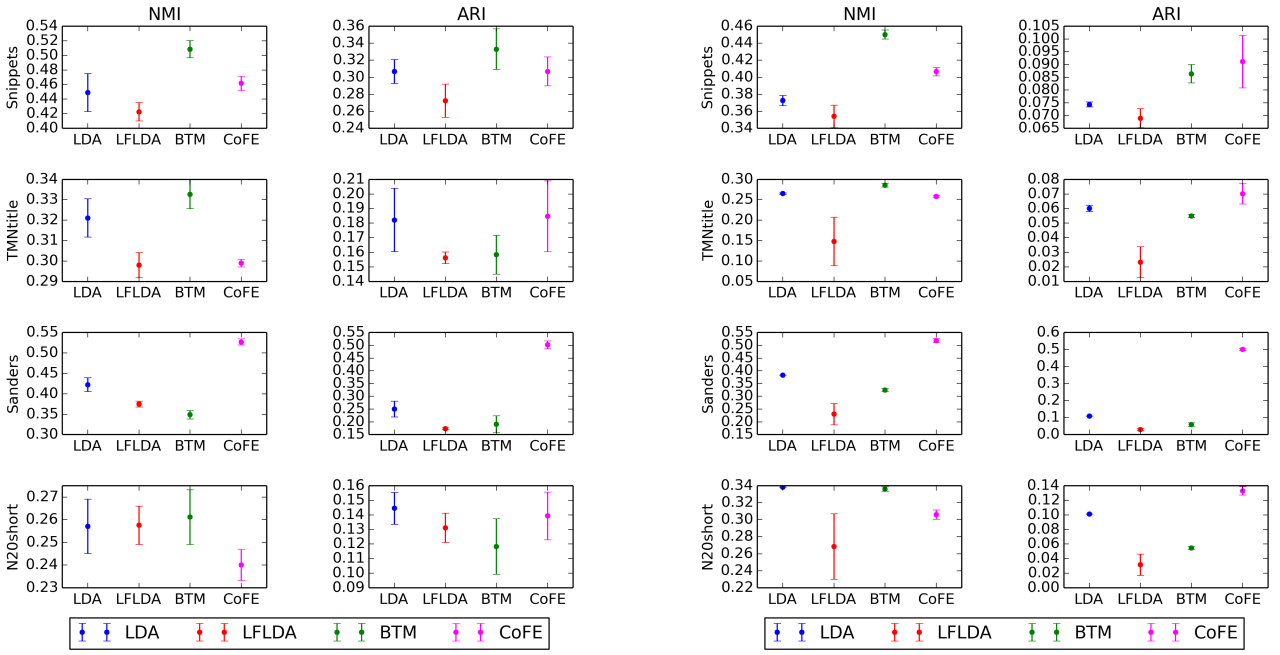
[1] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang, "Understanding the limiting factors of topic modeling via posterior contraction analysis," in *ICML*, 2014, pp. 190–198.

[3] 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*, 2011, pp. 338–349.

[4] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015.

[5] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *CIKM*, 2011, pp. 775–784.



(a) 20 Topics

(b) 100 Topics

Fig. 3. Clustering results for 20 and 100 topics. Error bars indicate the confidence interval.

TABLE V
TOPICS DISCOVERED BY LDA FROM THE ORIGINAL TNM AND ITS EXPANDED VERSION USING COFE.

	Original	CoFE	*sim*
1	case trial man police court charged guilty arrested charges death	case trial guilty murder man jury pleads accused charges court	0.86
2	apple google sony mobile ipad data deal facebook social million	apple google sony ipad mobile app video network social microsoft	0.85
3	billion report million banks ceo bid buy bank sec pay	billion report deal million buy ceo bid banks sec nyse	0.84
4	sales profit japan prices rise oil growth high year fall	prices stocks oil sales rise profit shares fed wall growth	0.83
5	china europe election obama debt vote party north korea south	europe china obama debt crisis imf election vote greek portugal	0.82
6	mets yankees win red rangers sox roundup game bruins canucks	yankees mets sox phillies win red indians marlins rays jays	0.81
7	nfl players draft lockout ncaa state judge nba court ohio	nfl players lockout draft goodell roger judge talks owners mediation	0.79
8	heat bulls knicks nba game lakers celtics thunder mavericks win	heat series game bulls lead roundup win nba celtics finals	0.78
9	bin laden pakistan killed police afghan kills death protest	pakistan nato laden bin afghan libya rebels kills attack	0.74
10	study drug law bill cancer texas house risk governor wisconsin	law bill governor union wisconsin texas passes house senate immigration	0.73
11	open final title nadal wins lead win state djokovic french	nadal final open djokovic french federer wins madrid cup beats	0.73
12	japan nuclear libya coast plant libyan rebels ivory quake nato	japan nuclear quake plant crisis radiation japanese tsunami plants disaster	0.71
13	review theater wedding dies royal man star war love kate	theater review idol critic american star corner love broadway jersey	0.67
14	libya yemen egypt middle israel east syria gaza lede bahrain	protesters yemen syrian forces protest protests syria middle bahrain fire	0.67
15	rail tornado south river derby storms texas dead tornadoes crash	tornado crash river south storms found dead space tornadoes town	0.66
16	coach news sheen space chicago charlie shuttle launch cooperative	news cooperative chicago show charlie talk sheen couric katie	0.53
17	study talk business time home kids focus world online green	study cancer risk drug linked heart diabetes drugs disease kids	0.43
18	sports briefing soccer world cup league dies times wins united	ncaa coach state sports basketball uconn tournament title butler college	0.39
19	critic corner health office recipes top idol box early star	dies royal wedding kate william film prince elizabeth middleton cannes	0.19
20	stocks street wall fed oil economic world japan bonds investors	rail derby morning line animal belmont kingdom favorite woods racing	0.07

[6] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proc. of the first Workshop on Social Media Analytics*. ACM, 2010, pp. 80–88.

[7] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *SIGIR*, 2013, pp. 889–892.

[8] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *WWW*, 2013, pp. 1445–1456.

[9] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *WWW*, 2008.

[10] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *ICLR*, 2013, pp. 1–12.

[11] D. Vitale, P. Ferragina, and U. Scaiella, "Classification of short texts by deploying topical annotations," in *Advances in Information Retrieval*. Springer, 2012.

[12] T. Minka, "Estimating a dirichlet distribution," MIT, Tech. Rep., 2000.

[13] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *GSCL*, pp. 31–40, 2009.

[14] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *EMNLP*, 2011, pp. 262–272.

[15] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *JCDL*, 2010, pp. 215–224.