Examining the Coherence of the Top Ranked Tweet Topics

Anjie Fang[1], Craig Macdonald[2], Iadh Ounis[2], Philip Habel[2]
[1]a.fang.1@research.gla.ac.uk, [2]{firstname.secondname}@glasgow.ac.uk
University of Glasgow, UK

ABSTRACT

Topic modelling approaches help scholars to examine the topics discussed in a corpus. Due to the popularity of Twitter, two distinct methods have been proposed to accommodate the brevity of tweets: the *tweet pooling method* and *Twitter LDA*. Both of these methods demonstrate a higher performance in producing more interpretable topics than the standard Latent Dirichlet Allocation (LDA) when applied on tweets. However, while various metrics have been proposed to estimate the coherence of the generated topics from tweets, the coherence of the top ranked topics, those that are most likely to be examined by users, has not been investigated. In addition, the effect of the number of generated topics K on the topic coherence scores has not been studied. In this paper, we conduct large-scale experiments using three topic modelling approaches over two Twitter datasets, and apply a state-of-the-art coherence metric to study the coherence of the top ranked topics and how K affects such coherence. Inspired by ranking metrics such as *precision at n*, we use *coherence at n* to assess the coherence of a topic model. To verify our results, we conduct a pairwise user study to obtain human preferences over topics. Our findings are threefold: we find evidence that *Twitter LDA* outperforms both LDA and the *tweet pooling method* because the top ranked topics it generates have more coherence; we demonstrate that a larger number of topics (K) helps to generate topics with more coherence; and finally, we show that *coherence at n* is more effective when evaluating the coherence of a topic model than the average coherence score.

1. INTRODUCTION

Topic modelling – e.g. Latent Dirichlet Allocation (LDA) [1] – is a widely used approach to discover latent topics within a corpus [4, 5]. As Twitter has gained in popularity, scholars have sought out ways to understand and model discussions on the forum. However tweet corpora are unlike other corpora (e.g. news articles and books), namely because they are short (limited to 140 characters), and they contain colloquial phrases and snippet text such as hashtags. Two well-known topic modelling applications for Twitter data are *Twitter LDA* (TLDA) [12] and LDA applied alongside the *tweet pooling method* (PLDA) [7]. Both methods have been shown to produce more interpretable topics than LDA on Twitter corpora [7, 12]. On the other hand, several metrics have been proposed in the literature to automatically estimate the coherence of the generated topic models. For example, Newman et al. [8] proposed a Pointwise Mutual Information(PMI)-based metric using Wikipedia as a background dataset to evaluate the coherence of a topic from news articles and books. More recently, a new coherence PMI-based metric using a Twitter background has been proposed for tweet corpora, and was found to be the closest to human judgements [2].

However, the coherence of the top ranked topics from tweets, those most likely to be examined by users, has not been previously investigated, nor has the effect of the number of generated topics K on the topic coherence scores. In this paper, we conduct large-scale experiments on two Twitter datasets to investigate the coherence of ranked topics generated by three topic modelling approaches (LDA, TLDA and PLDA). Inspired by the *precision at n* evaluation metric, we also explore the *coherence at n* scores of the generated topic models by using the state-of-the-art Twitter PMI-based coherence metric [2], which we describe in Section 3. The contributions of this paper are as follows: 1) we examine which of the three existing topic modelling approaches for Twitter data generates more coherent topics, 2) we analyse the relationship between the coherence of a topic model and the number of topics (K), and 3) we evaluate the utility of the *coherence at n* coherence metric for a topic model. To validate our findings, we perform a pairwise preference user study conducted on a crowdsourcing platform, where the workers are asked to choose the more coherent topic in a topic pair. The obtained human judgements are then compared with the coherence score of the generated topic models. We show, first, that TLDA performs better than LDA and PLDA; second, that a higher topic number K helps to produce topics with higher coherence; and finally, that the *coherence at n* scores are more effective than the commonly used averages in evaluating the coherence of a topic model.

2. BACKGROUND & RELATED WORK

LDA [1] models latent *topics* across *terms* and *documents*, where a topic is described as a probability distribution over terms Φ, and a document has its own probability distribution over topics. LDA has been used extensively to extract latent topics in standard corpora such as news articles and books [5]. Because tweets can be of lower quality due to the brevity of their content [6, 12], topics generated by using LDA are likely to be both mixed [11] and harder to inter-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914731

pret by the associated terms. In order to improve the performance of LDA for Twitter data, Mehrotra et al. [7] have proposed the *tweet pooling method*, which groups tweets into virtual documents (e.g. sharing the same author or the same hashtag). We refer to the application of LDA alongside the *tweet pooling method* as PLDA. TLDA is another approach from Zhao et al. [12], where a background term distribution is used to distinguish "real" topic terms from background terms, with the assumption that a single tweet contains a single topic. Indeed both PLDA and TLDA have outperformed LDA in producing more coherent topics [7, 12].

Newman et al. [8] conducted a user study to evaluate several proposed coherence metrics based on capturing the semantic similarity of the topics using external sources, e.g. Wikipedia and WordNet. They found that the most suitable coherence metric was based on estimating the Pointwise Mutual Information (PMI) of word pairs in a topic for corpora such as news articles. Stevens et al. [9] used the same Wikipedia PMI-based metric to analyse the coherence of the generated topic models from news articles.

Recently, a large-scale user study in [2] evaluated several coherence metrics including the Wikipedia PMI-based and WordNet-based metrics on tweets. Fang et al. [2] showed that a newly proposed coherence metric leveraging a Twitter background dataset, called the *Twitter PMI-based metric* (hereafter, T-PMI), has a markablely high agreement with human judgements on tweet corpora. In the following, we use the T-PMI metric to evaluate the coherence of the topics generated by using three topic modelling approaches: LDA, TLDA, and PLDA. We further study the coherence of the top ranked topics likely to be examined by users, and the effect of the number of generated topics K on the topic coherence scores. To the best of our knowledge, this paper contributes the first study of the coherence of the top ranked topics and how K affects such coherence on tweets.

3. METRICS

We now describe the T-PMI coherence metric, which analyses the coherence of topics and topic models. In the T-PMI metric [2], a topic t is first represented by the 10 most frequent words $\{w_1, w_2, ..., w_{10}\}$ selected by the topic term probabilities in Φ. Any two words among the 10 words in a topic make a word pair, e.g. Pair(w_i,w_j). The coherence C of a topic is measured by averaging the PMI score of all its word pairs using Equation (1), where we use the top 10 words representing a topic, i.e. $n = 10$. The PMI score of a word pair is pre-calculated using Equation (2) from a Twitter background dataset. We describe how we set up this metric in Section 4.

$$C(t) = \frac{1}{\sum_{m=1}^{n-1} m} \sum_{i=1}^{n} \sum_{j=i+1}^{n} PMI(w_i, w_j) \quad (1)$$

$$PMI(w_i, w_j) = log \frac{p(w_i, w_j)}{p(w_i) \times p(w_j)} \quad (2)$$

We use the three aforementioned topic modelling approaches to obtain the topic models from tweets, varying the topic number K. We use the T-PMI coherence metric to calculate the average coherence of a topic model, i.e. the coherence score across all K topics. Intuitively, the average coherence of a topic model reflects the quality of the entire topic model (containing K topics). However, end-users typically are only interested in the most coherent topics in a topic model, rather than the whole model. Inspired by ranking metrics such as the *precision at n* metric, we use *coherence at n* to evaluate the coherence of a topic model. In particular, *coherence at n* (coherence@n) indicates the average coherence score of the top n most coherent topics, where a topic is ranked by its coherence score. We argue that *coherence at n* can more effectively capture the coherence of a Twitter topic model.

Table 1: Two used Twitter datasets.

Name	Time Period	Users#	Tweets#
(1) MAY	1 to 31 May 2015	2,452	334,922
(2) TVD	8 to 10pm, 02 Apr 2015	121,594	343,511

4. DATASETS & EXPERIMENTAL SETUP

We use two datasets in our experiments. The first[1] is comprised of the tweets of 2,452 newspaper journalists in New York posted from 01/05/2015 to 31/05/2015, denoted here as MAY. We can reasonably assume that journalists discussed a large number of topics over this one-month period. The second dataset[1] consists of tweets related to the first TV debate among political party leaders during the UK General Election of 2015, denoted here as TVD. We expect that the number of topics covered over this short time period to be more limited, as conversations on Twitter tended to focus around the issues introduced during the two-hour debate. Details of these two datasets are shown in Table 1.

We use a Twitter background dataset from [2, 3], which contains 1%-5% random tweets posted from 01/01/ 2015 to 30/06/2015. Following [2, 3], we remove stopwords, terms occurring in less than 20 tweets and the retweets from this background dataset. The remaining tweets (30,151,847) are used to calculate the prior PMI score of the occurring word pairs in order to implement the T-PMI metric.

We use the Mallet[2] and Twitter LDA[3] toolkits to apply the three topic modelling approaches. For PLDA, we group the tweets posted by the same user in a given time interval into a virtual document[4]. The time interval is set to 10 minutes for TVD, and 6 hours for MAY, given the narrow time period of the TVD dataset and the more expansive one for the MAY dataset. The LDA parameters α and β are set to $50/K$ and 0.01 following [10], and the TLDA parameter γ is set to 20 following [12]. Since the TVD dataset contains just two hours of tweets, we set the maximum topic number K to 100, and then use 46 different K values between 10 and 100 (step = 2). We set K to a maximum of 500 for MAY, and use 49 different K values ranging from 10 to 500 (step = 10). Each topic modelling approach is run 5 times for each K. Thus, we obtain 5 topic models for each K. In the next section, we analyse the coherence of these 1,425 topic models ($46 \times 5 \times 3 + 49 \times 5 \times 3$).

5. COHERENCE OF TOPIC MODELS

Figure 1 shows the T-PMI coherence of three types of topic models (LDA, TLDA, and PLDA) for MAY and TVD, varying K. Each point in Figure 1 represents the average coherence or coherence@n score of 5 topics models, for $n = \{5, 10, 20, 30, 40\}$. For instance, the point (20, 0.0020) on the red line in Figure 1(a) shows that the average coherence score of the top 5 topics (coherence@5) in the 5 topic models where K=20 is 0.0020. Higher scores indicate better coherence.

First, it is clear that the average coherence (the solid black line) of all topics in a model decreases as K increases across

[1] Collected using the Twitter API. [2] mallet.cs.umass.edu
[3] github.com/minghui/Twitter-LDA [4] We do not group tweets by hashtags since there are not many hashtags in our datasets.

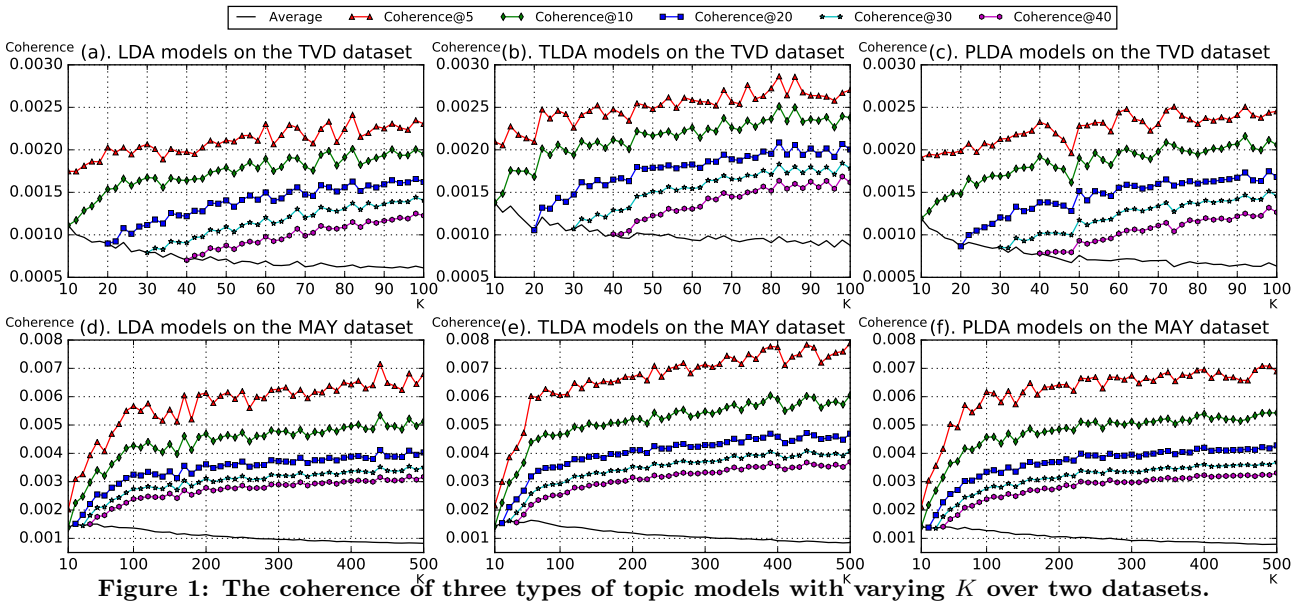

Figure 1: The coherence of three types of topic models with varying K over two datasets.

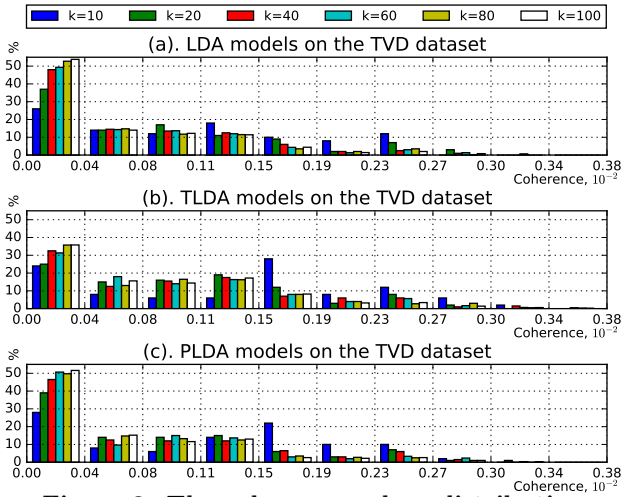

Figure 2: The coherence values distribution.

the three topic modelling approaches. These results are similar to what Stevens et al. [9] observed. However, the coherence@n score (represented by coloured lines with distinguishing symbols) of all topic models increases as K grows. This finding tells us that setting a higher number of topics, K, increases the coherence of the n most coherent topics. Second, the coherence@n score is higher for TLDA across the two datasets, which suggests that the top n topics in the TLDA models have a higher coherence. We also see that PLDA generates slightly more coherent topics than LDA, although the performance of PLDA is closer to LDA than to TLDA. Third, we observe that the average coherence and coherence@n score of the LDA & PLDA topic models on the TVD dataset become stable around K=80, while the coherence of the TLDA topic models on the MAY dataset has a local peak around K=390. It should be noted that the average coherence score cannot adequately capture the performance differences of the 3 topic modelling approaches on the MAY dataset (see Figure 1(d), (e) and (f)), and so coherence@n is preferred. Since a larger K leads to a higher computational cost, a K should be selected when the coherence of the topic model begins to stabilise or when it reaches a local peak.

Next, in Figure 2, we show the distribution of the topics' coherence scores for topic models with varying K on the TVD dataset. The coherence scores are distributed across 10 bins, to show the number of topics exhibiting different levels of coherence. First, the volume of topics with coherence $[0, 0.4^{-3})$ is highest across all topic models in the TVD dataset. As K increases, the topic models include more topics with less coherence. This is why the average coherence of the topic models declines as K increases, shown in Figure 1 by the solid black lines. Second, the volume of topics with coherence $[0, 0.4^{-3})$ is lowest in the TLDA models. This result further indicates that TLDA outperforms PLDA and LDA, as it generates fewer meaningless topics. We also observe the same pattern for the MAY dataset[5].

To verify that a larger K helps generate topics with higher coherence, and to examine the utility of the coherence@n metric, we conduct a user study in Section 6 where we compare human preferences with the topics' coherence scores from the T-PMI metric.

6. USER STUDY

Because it is challenging for humans to give a graded coherence score for topics, we conduct a pairwise preference user study. Similar studies have been conducted previously in [2, 3]. We recruited workers from CrowdFlower[6] and asked them to select the more coherent topic from two provided topics (a topic pair). We describe here how we generate the topic pairs, the CrowdFlower job, how we control the job quality, and the crowdsourcing results.

Generating Topic Pairs For a given approach (e.g. TLDA), we choose two $K = \{a,b\}$ $(a < b)$ values representing different topic model outcomes with difference coherences (recall that each approach is repeated 5 times, hence 10 models in total). From each selected topic model, we select the top n most coherent topics. Thus, we have two topic pools: $P_{k=a}$ & $P_{k=b}$, and each pool has $5 \times n$ topics[7]. To make the preference task easier for workers, we show two similar topics in a topic pair. First, we sample a number of topics from $P_{k=a}$ randomly. For each sampled topic ($t_j^{P_{k=a}}$), we use Equation (3) to identify its closest topic in $P_{k=b}$, where V_t is a vector representation using the term distribution of topic t. We denote the selected pairs of topics as Pairs($P_{k=a} \rightarrow P_{k=b}$). Likewise, we also generate the same number of Pairs($P_{k=b} \rightarrow P_{k=a}$). In our user study, we use TLDA topic models on the MAY dataset with K=$\{50,100,300,390\}$, since its K range is set wider

[5] The figure is not shown due to space limitations.
[6] crowdflower.com
[7] Each experiment is repeated 5 times.

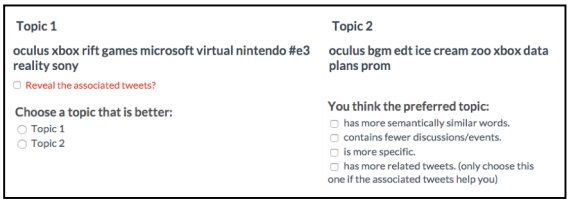

Figure 3: The CrowdFlower user interface.

than that of the TVD dataset. We compare the coherence of models with K=50 vs. K=300 (denoted as comparison Unit(50,300)) and topic models with K=100 vs. K=390 (comparison Unit(100,390)). Therefore, we can examine whether topic models with a larger K (300/390) have more coherent topics than the models with smaller K (50/100). For the comparisons Unit(50,300)/Unit(100,390), we select the top 30/20[8] topics for the topic pools ($P_{k=\{50,100,300,390\}}$). We generate 40 topic pairs for each comparison unit. Note that we ignore the topic pairs where two topics in the pair share the same top 10 words or share less than 3 mutual words among the top 10 words. Hence this ensures that the two selected topics in a topic pair are similar enough for humans.

$$closest(t_j^{P_{k=a}}) = argmax_{i<K}\ (cosine(V_{t_j^{P_{k=a}}}, V_{t_i^{P_{k=b}}})) \quad (3)$$

Job Description We present each CrowdFlower worker with the top 10 words (ranked by their probabilities in a topic) from two topics in a topic pair, labelled Topic 1 and Topic 2, along with their 3 most retweeted tweets[9]. Based on these 10 words, we ask the workers to select the more coherent topic among the two shown. We tell the worker that a more coherent topic is one that is less mixed and that can be interpreted. The workers are instructed to take into account: 1) the number of semantically similar words (e.g. President & Obama) among the 10 shown words, 2) whether the words shown suggest a mixed topic (i.e. more than one discussion), and 3) whether the words shown are more specific. A worker can also consider two associated tweets for the two topics if he/she cannot make a decision. We also offer guidance for using these tweets: 1) whether the 10 shown words are reflected by their tweets and 2) whether these tweets are related with the two topics. Figure 3 captures the user interface on CrowdFlower, along with an example of a topic pair. We collect a total of 5 judgements from 5 different workers for each topic pair. For each judgement, we paid workers $0.05.

Quality Control We used test questions for worker quality control. To set these test questions, we began by choosing a number of topic pairs, where the topic preference was manually verified in advance. Only those workers who passed the test were allowed to enter the task. The worker must have maintained more than 70% accuracy on the test questions throughout the whole task, otherwise their judgements were not used. We limited the workers' country to the United States as the MAY tweets were written by New York journalists. Our user study engaged 52 trusted workers.

Crowdsourcing Results Table 2 lists the human judgement results compared with the coherence scores from the T-PMI metric. For comparison unit (50,300) - Table 2(a) - the 40 topics we select from topic models with K=300 are significantly more coherent than those from topic models with K=50 according to both the human vote[10] fraction and the T-PMI coherence scores. We observe the same results for comparison unit (100,390), c.f. Table 2(b). This finding shows again that a larger K helps to obtain topics with more coherence. However, if one uses only the average coherence score to evaluate the topic model, then these models are indistinguishable. This finding suggests that the coherence@n metric should be used to measure the coherence quality of a topic model.

[8] The top 20 topics in Unit(100,390) are more distinguishable than the top 30. [9] These tweets help workers understand the topic. [10] A topic in a topic pair receives one vote when it is preferred by a human.

Table 2: The comparison of coherence scores.

(a). TLDA topic models with K=50 vs. K=300, Unit(50,300)

K	Human vote fraction	T-PMI
50	0.311	2.06^{-3}
300	$\mathbf{0.689}^{(*)}$	$\mathbf{3.15^{-3}}^{(*)}$

(b). TLDA topic models with K=100 vs. K=390, Unit(100,390)

K	Human vote fraction	T-PMI
100	0.411	3.54^{-3}
390	$\mathbf{0.589}^{(**)}$	$\mathbf{4.00^{-3}}^{(**)}$

$*/(**)$ denote $p < 0.01$ ($p < 0.05$) according to the Wilcoxon signed-rank test, compared to the smaller K.

7. CONCLUSIONS

This paper studied the coherence of Twitter topic models through a large scale experiment varying both the topic modelling approach and K, and thereafter verified conclusions using a pairwise user study. To summarise, we first found that *Twitter LDA* (TLDA) outperformed PLDA and LDA, as it generated topics with a higher coherence and also less meaningless topics. Second, we showed that increasing the number of topics (K) helped to generate topics with a higher coherence. Third, *coherence at n* is more effective in evaluating a topic model than the average coherence. Our paper has two implications for researchers implementing topic modelling on tweets: first, larger K values result in the top n topics being the most coherent ($n < K$); second, when evaluating the coherence of a topic model, scholars should use the *coherence at n* metric.

8. REFERENCES

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

[2] A. Fang, C. Macdonald, I. Ounis, and P. Habel. Topics in tweets: A user study of topic coherence metrics for Twitter data. In *Proc. of ECIR*, 2016.

[3] A. Fang, C. Macdonald, I. Ounis, and P. Habel. Using word embedding to evaluate the coherence of topics from Twitter data. In *Proc. of SIGIR*, 2016.

[4] A. Fang, I. Ounis, P. Habel, C. Macdonald, and N. Limsopatham. Topic-centric classification of Twitter user's political orientation. In *Proc. of SIGIR*, 2015.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of NAS*, 2004.

[6] L. Hong and B. D. Davison. Empirical study of topic modeling in Twitter. In *Proc. of SOMA*, 2010.

[7] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proc. of SIGIR*, 2013.

[8] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proc. of NAACL*, 2010.

[9] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. In *Proc. of EMNLP-CoNLL*, 2012.

[10] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427:424–440, 2007.

[11] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proc. of ICML*, 2014.

[12] 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 *Proc. of ECIR*, 2011.