Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data

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. Twitter offers scholars new ways to understand the dynamics of public opinion and social discussions. However, in order to understand such discussions, it is necessary to identify coherent topics that have been discussed in the tweets. To assess the coherence of topics, several automatic topic coherence metrics have been designed for classical document corpora. However, it is unclear how suitable these metrics are for topic models generated from Twitter datasets. In this paper, we use crowdsourcing to obtain pairwise user preferences of topical coherences and to determine how closely each of the metrics align with human preferences. Moreover, we propose two new automatic coherence metrics that use Twitter as a separate background dataset to measure the coherence of topics. We show that our proposed Pointwise Mutual Information-based metric provides the highest levels of agreement with human preferences of topic coherence over two Twitter datasets.

1 Introduction

Twitter is an important platform for users to express their ideas and preferences. In order to examine the information environment on Twitter, it is critical for scholars to understand the topics expressed by users. To do this, researchers have turned to topic modelling approaches [1, 2], such as Latent Dirichlet Allocation (LDA). In topic models, a document can belong to multiple topics, while a topic is considered a multinomial probability distribution over terms [3]. The examination of a topic's term distribution can help researchers to examine what the topic represents [4, 5]. To present researchers with interpretable and meaningful topics, several topic coherence metrics have been previously proposed [6–8]. However, these metrics were developed based on corpora of news articles and books, which are dissimilar to corpora of tweets, in that the latter are brief (i.e. < 140 characters), contain colloquial statements or snippets of conversation, and use peculiarities such as hashtags. Indeed, while topic modelling approaches specific to Twitter have been developed (e.g. Twitter LDA [2]), the suitability of these coherence metrics for Twitter data has not been tested.

In this paper, we empirically investigate the appropriateness of ten automatic topic coherence metrics, by comparing how closely they align with human judgments of topic coherence. Of these ten metrics, three examine the statistical coherence of a topic at the term/document distributions levels, while the

remaining seven consider if the terms within a topic exhibit semantic similarity, as measured by their alignment with external resources such as Wikipedia or WordNet. In this work, we propose two new coherence metrics based on semantic similarity, which use a separate background dataset of tweets.

To evaluate which coherence metrics most closely align with human judgments, we firstly use three different topic modelling approaches (namely LDA, Twitter LDA (TLDA) [2], and Pachinko Allocation Model (PAM) [9]) to generate topics on corpora of tweets. Then, for pairs of topics, we ask crowdsourcing workers to choose what they perceive to be the more coherent topic. By considering the pairwise preferences of the workers, we then identify the coherence metric that is best aligned with human judgments.

Our contributions are as follows: 1) we conduct a large-scale empirical crowdsourced user study to identify the coherence of topics generated by three different topic modelling approaches upon two Twitter datasets; 2) we use these pairwise coherence preferences to assess the suitability of 10 topic coherence metrics for Twitter data; 3) we propose two new topic coherence metrics, and show that our proposed coherence metric based on Pointwise Mutual Information using a Twitter background dataset is the most similar to human judgments.

The remainder of this paper is structured as follows: Section 2 provides an introduction to topic modelling; Section 3 reports the related work of evaluating topic models; Section 4 describes 10 topic coherence metrics; Section 5 shows how we compare automatic metrics to human judgments; Section 6 describes the Twitter datasets we use in the user study (Section 7), while the experimental setup and the results are discussed in Section 8 and Section 9. Finally, we provide concluding summaries in Section 10.

2 Background: Topic Modelling

Topic modelling approaches can be used to identify coherent topics of conversation in social media such as Twitter [1, 2]. However, ensuring that the topic modelling approaches obtain coherent topics from tweets is challenging. Variants of LDA have been proposed to improve the coherence of the topics, while automatic metrics of topical coherence have also been proposed (see Section 3). However, as we argue in Section 3, the suitability of the automatic coherence metrics has not been demonstrated on Twitter data.

LDA [10] is one of the most popular topic modelling approaches. TLDA and PAM are two extensions of LDA. LDA is a Bayesian probabilistic topic modelling approach, where K latent topics (z) are identified, which are associated to both documents and terms, denoted as $P(z|d)$ and $P(w|z)$, respectively. PAM [9] is a 4-level hierarchical extension of LDA, where a document is represented by a multinomial distribution over super-topics θr, where a super-topic is a multinomial distribution θt over sub-topics. This structure helps to capture the relation between super-topics and sub-topics. PAM generates topics with higher coherence, improving the likelihood of held-out documents and improving the accuracy of classification [11]. On the other hand, Zhao et al. [2] recognised that due

to their brevity, tweets can be challenging for obtaining coherent topic models. To counter this, they proposed TLDA, which employs a background Bernoulli term distribution, where a Bernoulli distribution π controls the selection between "real" topic terms and background terms. Moreover, Zhao et al. [2] assumed that a single tweet contained a single topic. Based on human judgments, they showed that TLDA outperformed the standard LDA for discovering topics in tweets. Indeed, both TLDA and PAM have been reported to produce more coherent topics than LDA. Hence, we apply the three aforementioned approaches to extract topics from Twitter corpora. In the following section, we review various topic coherence metrics.

3 Related Work: Evaluating Topic Models

The early work on evaluating topic models calculated the likelihood of held-out documents [12]. Chang et al. [13] deployed a user study for the interpretation of the generated topics, by comparing human judgments to the likelihood-based measures. However, it was shown that a model that had a good held-out likelihood performance can still generate uninterpretable topics.

Mei et al. [4] provided a method to interpret the topics from topic models. Their approach relied on the statistical analysis of a topic's term distribution. Similarly, AlSumait et al. [6] used another statistical analysis metric to evaluate the topics. In this paper, we compare their metrics to human judgments that assess the coherence of topics. Newman et al. [7, 8] offered another way to evaluate the coherence of topics. They captured the semantically similar words among the top 10 terms in a topic and calculated the semantic similarity of the words using external resources, e.g. WordNet [14] and Wikipedia. They showed that the evaluation metric based on the Pointwise Mutual Information estimate of the word pairs generated from Wikipedia was the closest to human judgments.

The datasets used in [6–8] consisted of news articles and books; however Twitter data is different from the classical text corpora. Therefore, it is unclear how well these evaluation metrics perform when measuring the coherence of a topic in tweets. In the next section, we give more details about these metrics and our proposed new ones.

4 Automatic Topic Coherence Metrics

In this section, we describe the topic coherence metrics that we use to automatically evaluate the topics generated by topic modelling approaches. There are two types of coherence metrics: 1) metrics based on semantic similarity (introduced in [7, 8]) and 2) metrics based on statistical analysis (introduced in [6]). We propose two new metrics based on semantic similarity, which use a Twitter background dataset.

4.1 Metrics based on Semantic Similarity

In metrics based on semantic similarity, a topic is represented by the top 10 words ($\{w_1, w_2, ..., w_{10}\}$) ranked according to its term probabilities ($p(w|z)$) in

the term distribution ϕ. A word pair of a topic is composed by any two words from the topic's top 10 words. The coherence of a topic is measured by averaging the semantic similarities of all word pairs [7, 8] shown in Equation (1) below. In this paper, the *Semantic Similarity SS* of a word pair is computed by using three external resources: WordNet, Wikipedia and a Twitter background dataset.

$$Coherence(topic) = \frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10} SS(w_i, w_j) \quad (1) \qquad PMI(w_i, w_j) = log\frac{p(w_i, w_j)}{p(w_i), p(w_j)} \quad (2)$$

WordNet. WordNet groups words into synsets [14]. There are a number of semantic similarity and relatedness methods in the existing literature. Among them, the method designed by Leacock et al. [15] (denoted as LCH) and that designed by Jiang et al. [16] (denoted as JCN) are especially useful for discovering lexical similarity [8]. Apart from these two methods, Newman et al. [8] also showed that the method from Lesk et al. [17] (denoted as LESK) performs well in capturing the similarity of word pairs. Therefore, we select these 3 WordNet-based methods to calculate the semantic similarities of the topic's word pairs, and produce a topic coherence score.

Wikipedia. Wikipedia has been previously used as background data to calculate the semantic similarity of words [18, 19]. In this paper, we select two popular approaches in the existing literature on calculating the semantic similarity of words: *Pointwise Mutual Information* (PMI) and *Latent Semantic Analysis* [20] (LSA). PMI is a popular method to capture semantic similarity [7, 8, 18]. Newman et al. [7, 8] reported that the performance of PMI was close to human judgments when assessing the topic's coherence. Here the PMI data (denoted as W-PMI) is computed by using Equation (2) consisting of the PMI score of word pairs from Wikipedia. On the other hand, since it has been reported that the performance of PMI is no better than LSA on capturing the semantic similarity of word pairs [19], in this paper we also use LSA to obtain the similarity of the word pairs. In the LSA model, a corpus is represented by a term-document matrix. The cells represent the frequency of a term occurring in a document. To reduce the dimension of this matrix, a *Singular Value Decomposition* is applied on the matrix using the k largest singular values. After the decomposition, each term is represented by a dense vector in the reduced LSA space. The semantic similarity of terms can be computed by the distance metrics (e.g. cosine similarity) between the terms' vectors. We use Wikipedia articles as background data and calculate the LSA space (denoted as W-LSA), which is a collection of term vectors in 300 dimensions described in [21].

Twitter Background Dataset. Since tweets contain abbreviations and hashtags[1], Wikipedia cannot capture their semantic similarity. Hence, we crawl an additional Twitter background corpus of 1%-5% random tweets from 1 Jan 2015 to 30 June 2015 on Twitter. The background collection is likely to better reflect the semantic similarity of words that occur on Twitter. We use the same method as for Wikipedia to obtain our proposed two new metrics: the reduced LSA space (300 dimensions, denoted as T-LSA) and the PMI score of word pairs (denoted as T-PMI) that appear in each tweet.

[1] Note that many hashtags are not recorded in Wikipedia.

4.2 Metrics based on Statistical Analysis

Properties of how the term or documents are assigned to the topics can be indicative of the coherence of a topic model. In this section, we describe the term/document distributions of 3 types of meaningless topics defined in [6]: a uniform distribution over terms; a semantically vacuous distribution over terms; and a background distribution over documents. We explain how these permit the measurement of the coherence of a topic.

Uniform Term Distribution. In a topic's term distribution, if all terms tend to have an equal and constant probability, this topic is unlikely to be meaningful nor easily interpreted. A typical uniform term distribution ϕ_{uni} is defined in Equation (3), where i is the term index and N^k is the total number of terms in topic k.

$$\phi_{uni} = \{P(w_1), P(w_2), ..., P(w_{N^k})\},\ P(w_i) = \frac{1}{N^k} \tag{3}$$

Vacuous Term Distribution. A "real" topic should contain a unique collection of highly used words distinguishing this topic from the other topics. A topic is less coherent if a topic is mixed. A vacuous term distribution θ_{vac} represents a mixed term distribution, in which the term probability reflects the frequency of the term in the whole corpus. ϕ_{vac} is defined by Equation (4), where d is the document index and D is the total number of documents.

$$\phi_{vac} = \{P(w_1), P(w_2), ..., P(w_{N^k})\},\ P(w_i) = \sum_{k=1}^{K} \phi_{i,k} \times \frac{\sum_{d=1}^{D} \theta_{d,k}}{D} \tag{4}$$

Background Document Distribution. A "real" topic should represent documents within a semantically coherent theme. If a topic is close to most of the documents in the corpus, it is likely to be less meaningful and less coherent. Whereas the previous two distributions use terms to define the incoherent distribution of a topic, the topic distribution over documents can also reflect the quality of the topic [6]. A topic's document distribution ϑ^k is defined in Equation (5) and a typical background document distribution ϑ_{gb} is defined in Equation (6).

$$\vartheta^k = \{P(z = k|d_1), P(z = k|d_2), ..., P(z = k|d_D)\} \tag{5}$$

$$\vartheta_{gb} = \{P(d_1), P(d_2), ..., P(d_D)\},\ P(d_i) = \frac{1}{D} \tag{6}$$

Given a topic k, the coherence of the topic is calculated by measuring the Kullback Leibler divergence between this topic and those three meaningless topics described above. A small divergence indicates that the topic is less coherent. Hereafter, we use U (uniform), V (vacuous) and B (background) to denoted three metrics corresponding to the coherence functions $Coherence^U(k)$, $Coherence^V(k)$ and $Coherence^B(k)$ in Equation (7), respectively.

$$\begin{gathered} Coherence^U(k) = KL(\phi_{uni}||\phi^k), Coherence^V(k) = KL(\phi_{vac}||\phi^k) \\ Coherence^B(k) = KL(\vartheta_{gb}||\vartheta^k) \end{gathered} \tag{7}$$

In summary, in this paper we describe 7 metrics based on semantic similarity: LCH, JCN, LESK, W-LSA, W-PMI, T-LSA & T-PMI, and 3 metrics based on

Table 1. Comparison Task.

Comparison Unit	Topic Pairs in Unit
(1) Unit(LDA, TLDA)	Pairs(LDA→TLDA & TLDA→LDA)
(2) Unit(LDA, PAM)	Pairs(LDA→PAM & PAM→LDA)
(3) Unit(TLDA, PAM)	Pairs(TLDA→PAM & PAM→TLDA)

the statistical analysis of term/document distributions: U, V & B. Among them, T-LSA & T-PMI are our newly proposed metrics. In the following section, we present our approach to compare the discussed automatic coherence metrics to human judgments when assessing the coherence of topics.

5 Comparison of Coherence Metrics

In this section, we describe the methodology we use to identify whether the topic coherence metrics are aligned with human evaluations of topic coherence. It can be a challenging task for humans to produce graded coherence assessments of topics. Therefore, we apply a pairwise preference user study [22] to gather human judgments. A similar method has been previously used to compare summarisation algorithms [23]. In the rest of this section, we describe this comparison method.

Generating Topic Pairs. To compare the three topic modelling approaches, we divide the comparison task into three units: LDA vs. TLDA, LDA vs. PAM and TLDA vs. PAM. Each comparison unit consists of a certain number of topic pairs and each pair contains a topic from topic models T_1 and T_2, respectively (e.g. LDA vs. TLDA). To make the comparisons easier for humans, we present similar topics in a pair. Specifically, each topic model has a set of candidate topics, and each topic is represented as a topic vector using its term distribution. First, we randomly select a certain number of topics from topic model T_1. For each topic selected in T_1, we use Equation (8) to select the closest topic in T_2 using cosine similarity. The selected topic pairs are denoted as Pairs($T_1 \rightarrow T_2$). Moreover, we also generate the same number of topic pairs Pairs($T_2 \rightarrow T_1$) for comparison unit(T_1,T_2). Therefore, every comparison unit has a set of topic pairs shown in Table 1.

$$closest(topic_j^{T_1}) = argmin_{i<K}\ (1 - cosine(Vector_{topic_j^{T_1}}, Vector_{topic_i^{T_2}})) \quad (8)$$

Automatic Topic Coherence Evaluation. We use the topic coherence metrics described in Section 4 to rank the three topic modelling approaches: LDA, TLDA and PAM. For each topic in each topic pair, an automatic coherence metric gives a coherence score to each topic respectively. Thus, for each comparison unit, there is a group of data pairs. We apply the Wilcoxon signed-rank test to calculate the significance level of the difference between the two groups of data sample. For each comparison unit, an automatic coherence metric determines the better topic model between two approaches (e.g. LDA > TLDA), which results in a ranking order of the three topic modelling approaches. For instance, given the preferences LDA>TLDA, LDA>PAM & TLDA>PAM, we can obtain the

Table 2. The details of the two used Twitter datasets.

Name	Time Period	# of Users	# of Tweets
(1) NYJ	20/05/2015-19/08/2015	2,853	946,006
(2) TV debate	8pm-10pm 02/04/2015	121,594	343,511

ranking order LDA(1^{st})>TLDA(2^{nd})>PAM(3^{rd}). However, while it is possible for the preference results of comparison units not to permit a ranking order to be obtained – i.e. a Condorcet paradox such as TLDA>LDA, LDA>PAM & PAM>TLDA – we did not observe any such paradoxes in our experiments.

Human evaluation. Similarly as above, we also rank the three topic modelling approaches using the topic coherence assessments from humans described in Section 7. This obtained ranking order generated from humans is compared to that generated from the ten automatic coherence metrics to ascertain the most suitable coherence metric when assessing a topic's coherence.

6 Twitter Datasets

In our experiments, we use two Twitter datasets to compare the topic coherence metrics. The first dataset we use consists of tweets posted by 2,853 newspaper journalists in the state of New York from 20 May 2015 to 19 Aug 2015, denoted as NYJ. To construct this dataset, we tracked the journalists' Twitter handles using the Twitter Streaming API[2]. We choose this dataset due to the high volume of topics discussed by journalists on Twitter. The second dataset contains tweets related to the first TV debate during the UK General Election 2015. This dataset was collected by searching the TV debate-related hashtags and keywords (e.g. #TVDebate and #LeaderDebate) using the Twitter Streaming API. We choose this dataset because social scientists want to understand what topics people discuss. Table 2 reports the details of these two datasets. We describe our user study and experimental setups in Section 7 and Section 8, respectively.

7 User Preferences Study

In this section, we describe the method we use to obtain the human ground-truth ranking order of the three topic modelling approaches. As described in Section 5, the comparison task is divided into three comparison units. Each comparison unit has two sets of topic pairs from the NYJ and TV debate datasets respectively. We asked humans to conduct a pairwise preference evaluation, and we then used the obtained human' preferences of topics from the topic models to rank the three topic modelling approaches. For collecting human judgments, we used the CrowdFlower[3] crowdsourcing platform.

CrowdFlower Job Description. For each topic pair in our three comparison units, we present a worker (i.e. a human) with the top 10 highly frequent words from the two topics (a topic pair, generated from two topic modelling approaches) along with their associated 3 most retweeted tweets, which are likely to

[2] dev.twitter.com [3] crowdflower.com

Fig. 1. The designed user interface and the associated tweets for two topics.

represent the topic. A CrowdFlower worker is asked to choose the more coherent topic from two topics using these 10 words. To help the workers understand and finish the task, we provide guidelines that define a more coherent topic as one that contains fewer discussions/events and that can be interpreted easily. We instruct workers to consider: 1) the number of semantically similar words among the 10 shown words, 2) whether the 10 shown words imply multiple topics and 3) whether the 10 shown words have more details about a discussion/event. If a decision cannot be made with these 10 words, a worker can then use the optional 3 associated tweets, shown in Figure 1. We provide two guidelines for using these tweets for assistance: 1) consider the number of the 10 shown words from a topic that can be reflected by the tweets and 2) consider the number of tweets that are related with the topic. After the workers make their choices, they are asked to select the reasons, as shown in Figure 1. The CrowdFlower workers were paid \$0.05 for each judgment per topic pair. We gather 5 judgments for each topic pair from 5 different workers.

CrowdFlower Quality Control. To ensure the quality of the CrowdFlower judgments, we use several quality control strategies. First, we provide a set of test questions, where for each question workers are asked to choose a topic preference from a topic pair. The answers of the test questions are verified in advance. Only workers that pass the test are allowed to enter the task. Moreover, the worker must have maintained 70% or more accuracy on the test questions in the task, otherwise their judgments are erased. Since the NYJ dataset is related to the United States, we limit the workers country to the United States only. The TV debate dataset contains topics that can be easily understood, and thus we set the workers country to English speaking countries (e.g. United Kingdom, United States, etc.). Overall, 77 different trusted workers for the NYJ dataset and 91 for the TV debate dataset were selected, respectively.

Human Ground-truth Ranking Order. As described above, we obtain 5 human judgments for each topic pair. A topic receives one vote if it is preferred by one worker. Thus, we assign each topic in each topic pair a fraction of the 5 votes received. A higher number of votes indicates that the topic is judged as being more coherent. Hence, for each comparison unit, we obtain a number of data pairs. Then, we apply the methodology described in Section 5 to obtain the human ground-truth ranking order of the three topic modelling approaches, i.e. 1^{st}, 2^{nd}, 3^{rd}.

Table 3. The size of LSA space and the number of word pairs.

Model	Original Size of Matrix # of term × # of Doc	Model	# of word pairs
(1) W-LSA	1,096,192 × 3,873,895	(1)W-PMI	179,110,791
(2) T-LSA	609,878 × 30,151,847	(2) T-PMI	354,337,473

8 Experimental Setup

In this section, we describe the experimental setup for generating the topics and implementing the automatic metrics.

Generating topics. We use Mallet[4] and Twitter LDA[5] to deploy the three topic modelling approaches on the two datasets (described in Section 6). The LDA hyper-parameters α and β are set to $50/K$ and 0.01 respectively, which work well for most corpora [3]. In TLDA, we follow [2] and set γ to 20. We set the number of topics K to a higher number, 100, for the NYJ dataset as it contains many topics. The TV debate dataset contains fewer topics, particularly as it took place only over a 2 hour period, and politicians were asked to respond to questions on specific themes and ideas[6]. Hence, we set K to 30 for the TV debate dataset. Each topic modelling approach is run 5 times for the two datasets. Therefore, for each topic modelling approach, we obtain 500 topics in the NYJ dataset and 150 topics in the TV debate dataset. We use the methodology described in Section 5 to generate 100 topic pairs for each comparison unit. For example, for comparison Unit(LDA,TLDA), we generate 50 topic pairs of Pairs(LDA→TLDA) and 50 topic pairs of Pairs(TLDA→LDA).

Metrics Setup. Our metrics using WordNet (LCH, JCN & LESK) are implemented using the *WordNet::Similarity* package. We use the Wikipedia LSA space and the PMI data from the *SEMILAR* platform[7] to implement the W-LSA and W-PMI metrics. Since there are too many terms and tweets in our Twitter background dataset, we remove stopwords, terms occurring in less than 20 tweets, tweets with less than 10 terms and retweeting tweets. These steps help to reduce the computational complexity of LSA and PMI using this Twitter background dataset. After this pre-processing, the number of remaining tweets is 30,151,847. Tables 3 shows the size of T-LSA space and the number of word pairs in T-PMI.

9 Results

We first compare the ranking order of the three topic modelling approaches using the automatic coherence metrics and human judgments. Then we show the differences between each of the automatic metric and human judgments.

Table 4 reports the average coherence score of the three topic models using the ten automatic metrics (displayed in white background). We also average the fraction of human votes of the three topic models, shown in Table 4 as column "human"(shown in grey background). We apply the methodology introduced

[4] mallet.cs.umass.edu [5] github.com/minghui/Twitter-LDA [6] goo.gl/JtzJDz
[7] semanticsimilarity.org

Table 4. The results of the automatic topic coherence metrics on the two datasets and the corresponding ranking orders. "×" means no statistically significant differences ($p \leq 0.05$) among the three topic modelling approaches. Two topic modelling approaches have the same rank if there are no significant differences between them.

NYJ												
	LCH	Rank	JCN	Rank	LESK	Rank	W-LSA	Rank	W-PMI	Rank	T-LSA	Rank
LDA	0.517		0.020		0.028		0.157	$1^{st}/2^{nd}$	0.205	1^{st}	0.014	
TLDA	0.494	×	0.019	×	0.018	×	0.132	$1^{st}/2^{nd}$	0.190	2^{nd}	0.004	×
PAM	0.544		0.021		0.009		0.073	3^{rd}	0.150	3^{rd}	0.011	
	T-PMI	Rank	U	Rank	V	Rank	B	Rank	Human	Rank		
LDA	1.63e-3	1^{st}	0.092		0.548		1.365	1^{st}	0.636	1^{st}		
TLDA	1.52e-3	2^{nd}	0.196	×	0.529	×	0.828	2^{nd}	0.553	2^{nd}		
PAM	4.53e-4	3^{rd}	-0.074		0.542		-3.473	3^{rd}	0.129	3^{rd}		

TV debate												
	LCH	Rank	JCN	Rank	LESK	Rank	W-LSA	Rank	W-PMI	Rank	T-LSA	Rank
LDA	0.448		0.017		0.014		-0.019	$2^{nd}/3^{rd}$	0.134	$1^{st}/2^{nd}$	-0.033	
TLDA	0.434	×	0.016	×	0.014	×	0.064	1^{st}	0.141	$1^{st}/2^{nd}$	-0.019	×
PAM	0.502		0.020		0.016		-0.041	$2^{nd}/3^{rd}$	0.127	3^{rd}	-0.023	
	T-PMI	Rank	U	Rank	V	Rank	B	Rank	Human	Rank		
LDA	3.57e-4	$2^{nd}/3^{rd}$	0.293	$1^{st}/2^{nd}$	0.548		-1.31	$1^{st}/2^{nd}$	0.475	$2^{nd}/3^{rd}$		
TLDA	4.11e-4	1^{st}	0.248	3^{rd}	0.535	×	-0.606	$1^{st}/2^{nd}$	0.590	1^{st}		
PAM	3.26e-4	$2^{nd}/3^{rd}$	0.304	$1^{st}/2^{nd}$	0.515		-2.092	3^{rd}	0.431	$2^{nd}/3^{rd}$		

in Section 5 to obtain the ranking orders shown in Table 4 as column "rank". By comparing the human ground-truth ranking orders of the three topic modelling approaches, we observe that the three topic modelling approaches perform differently over the two datasets.

Firstly, we observe that the ranking order from our proposed PMI-based metric using the Twitter background dataset (T-PMI) best matches the human ground-truth ranking order across our two Twitter datasets. This indicates that T-PMI can best capture the performance differences of the three topic modelling approaches. However, our other proposed metric T-LSA does not allow statistically distinguishable differences between topic modelling approaches to be identified (denoted by "×"). Second, for metrics based on semantic similarity, both W-PMI and W-LSA produce the same or a similar[8] ranking order as humans on the two datasets. However, both W-PMI and W-LSA perform no better than T-PMI metric. On the other hand, for metrics based on statistical analysis, the B metric (statistical analysis on the document distribution) can also lead to a similar performance as W-LSA or W-PMI compared to human judgments. Moreover, our results show that the remaining metrics perform no better than T-PMI, W-PMI & W-LSA metrics according to the ranking orders, i.e. their ranking orders do not match the human ground-truth ranking order.

To further compare the automatic coherence metrics and human judgments, we use the sign test to determine whether the 10 automatic metrics perform differently than human judgments. Specifically, for an automatic metric or human judgments, we obtain 100 preference data points from 100 topic pairs for a comparison unit(e.g. Unit(T_1,T_2)), where "1"/"-1" represents that the topic from T_1/T_2 is preferred and "0" means no preference. Then, we hypothesise that there are no differences between the preference data points from an automatic

[8] Part of the order matches the order from humans.

Table 5. The obtained p-values from the sign tests.

NYJ										
	LCH	JCN	LESK	W-LSA	W-PMI	T-LSA	T-PMI	U	V	B
LDA vs. TLDA	0.104	0.133	**0.039**	0.783	0.779	0.097	0.410	**4.1e-11**	0.787	**2.2e-13**
TLDA vs. PAM	**2.7e-9**	**3.8e-10**	**0.0**	**1.8e-7**	**1.1e-4**	**1.7e-10**	1.0	**0.007**	**8.1e-13**	**0.007**
LDA vs. PAM	**2.2e-13**	**3.4e-11**	**7.2e-14**	**0.001**	0.210	**3.0e-11**	0.145	1.0	**2.4e-10**	**0.003**
TV debate										
	LCH	JCN	LESK	W-LSA	W-PMI	T-LSA	T-PMI	U	V	B
LDA vs. TLDA	**0.010**	0.104	0.075	0.999	0.401	0.651	0.999	**1.2e-6**	**2.0e-5**	**0.011**
TLDA vs. PAM	**0.003**	**0.007**	**0.005**	0.211	0.568	**0.010**	0.783	**4.7e-5**	**0.003**	**3.6e-12**
LDA vs. PAM	0.174	**0.007**	0.576	0.671	0.391	0.791	0.882	0.391	0.895	0.202

metric and that from humans for a comparison unit (null hypothesis), and thus we calculate the p-values reported in Table 5. Each metric gets 6 tests (3 tests from the NYJ dataset and 3 tests from the TV debate dataset). If $p \leq 0.05$, the null hypothesis is rejected, which means that there are differences between the preferences of the same comparison unit between a given metric and humans.

We observe that the null hypotheses of 6 tests of T-PMI metric are not rejected across the two datasets. This suggests that T-PMI is the most aligned coherence metric with human judgments since there are no differences between T-PMI and human judgments for all the comparison units (shown in Table 5, $p \geq 0.05$). Moreover, only one test of W-PMI shows preference differences in a comparison unit (i.e. Unit(TLDA,PAM) in the NYJ dataset, where the null hypothesis is rejected) while W-LSA gets two tests rejected. Apart from these three metrics, the tests of the other metrics indicate that there are significant differences between these metrics and human judgments in most of comparison units. In summary, we find that the T-PMI metric demonstrates the best alignment with human preferences.

10 Conclusions

In this paper, we used three topic modelling approaches to evaluate the effectiveness of ten automatic topic coherence metrics for assessing the coherence of topic models generated from two Twitter datasets. Moreover, we proposed two new topic coherence metrics that use a separate Twitter dataset as background data when measuring the coherence of topics. By using crowdsourcing to obtain pairwise user preferences of topical coherences, we determined how closely each of the ten metrics align with the human judgments. We showed that our proposed PMI-based metric (T-PMI) provided the highest levels of agreement with the human assessments of topic coherence. Therefore, we recommend its use in assessing the coherence of topics generated from Twitter. If Twitter background data is not available, then we suggest one use PMI-based and LSA-based metrics using Wikipedia as a background (c.f. W-PMI & W-LSA). Among the metrics not requiring background data, the B metric (statistical analysis on the document distribution) is the most aligned with user preferences. For future work, we will investigate how to use the topic coherence metrics such that the topic modelling approaches can be automatically tuned to generate topics with high coherence.

References

1. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. In: Proc. of SOMA. (2010)
2. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: Proc. of ECIR. (2011)
3. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of latent semantic analysis **427**(7) (2007) 424–440
4. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proc. of SIGKDD. (2007)
5. Fang, A., Ounis, I., Habel, P., Macdonald, C., Limsopatham, N.: Topic-centric classification of Twitter user's political orientation. In: Proc. of SIGIR. (2015)
6. AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic significance ranking of lda generative models. In: Proc. of ECMLPKDD. (2009)
7. Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. In: Proc. of ADCS. (2009)
8. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Proc. of NAACL. (2010)
9. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proc. of ICML. (2006)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3** (2003) 993–1022
11. Li, W., Blei, D., McCallum, A.: Nonparametric bayes pachinko allocation. In: Proc. of UAI. (2007)
12. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proc. of ICML. (2009)
13. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proc. of NIPS. (2009)
14. Fellbaum, C.: WordNet. Wiley Online Library (1998)
15. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database **49**(2) (1998) 265–283
16. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of ICRCL. (1997)
17. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proc. of SIGDOC. (1986)
18. Rus, V., Lintean, M.C., Banjade, R., Niraula, N.B., Stefanescu, D.: SEMILAR: The semantic similarity toolkit. In: Proc. of ACL. (2013)
19. Recchia, G., Jones, M.N.: More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. Behavior research methods **41**(3) (2009) 647–656
20. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes **25**(2-3) (1998) 259–284
21. Stefănescu, D., Banjade, R., Rus, V.: Latent semantic analysis models on wikipedia and TASA. In: Proc. of LREC. (2014)
22. Carterette, B., Bennett, P.N., Chickering, D.M., Dumais, S.T.: Here or there. In: Proc. of ECIR. (2008)
23. Mackie, S., McCreadie, R., Macdonald, C., Ounis, I.: On choosing an effective automatic evaluation metric for microblog summarisation. In: Proc. of IIiX. (2014)