Text, Topics, and Turkers: A Consensus Measure for Statistical Topics

Fred Morstatter
Arizona State University
699 S. Mill Ave.
Tempe, AZ, 85281
fred.morstatter@asu.edu

Jürgen Pfeffer
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA, 15213
jpfeffer@cs.cmu.edu

Katja Mayer
University of Vienna
Universitätsring 1
1010 Vienna, Austria
katja.mayer@univie.ac.at

Huan Liu
Arizona State University
699 S. Mill Ave.
Tempe, AZ 85281
huanliu@asu.edu

ABSTRACT

Topic modeling is an important tool in social media analysis, allowing researchers to quickly understand large text corpora by investigating the topics underlying them. One of the fundamental problems of topic models lies in how to assess the quality of the topics from the perspective of human interpretability. How well can humans understand the meaning of topics generated by statistical topic modeling algorithms? In this work we advance the study of this question by introducing *Topic Consensus*: a new measure that calculates the quality of a topic through investigating its consensus with some known topics underlying the data. We view the quality of the topics from three perspectives: 1) topic interpretability, 2) how documents relate to the underlying topics, and 3) how interpretable the topics are when the corpus has an underlying categorization. We provide insights into how well the results of Mechanical Turk match automated methods for calculating topic quality. The probability distribution of the words in the topic best fit the Topic Coherence measure, in terms of both correlation as well as finding the best topics.

Categories and Subject Descriptors

I.5.4 [Applications]: Text Processing; I.2.7 [Natural Language Processing]: Language Generation

Keywords

Topic Analysis; Topic Modeling; Text Analysis; Text Mining

INTRODUCTION

Text analysis has proven to be one of the cornerstones of social media analysis. Researchers can usually rely on text to give them signal for their specific problem. Text has been used to greatly aid problems such as opinion mining [27], and user home location detection [15], and to find users in crisis situations [19]. Topic modeling is one of the fundamental text analysis techniques in social media. Topic modeling takes a corpus of text and finds the underlying topics in it. This process is akin to organizing newspaper articles by the “section” in which they appear, and simultaneously ranking words for that section. Topic modeling algorithms have been widely used for many tasks in social media research, such as using text to find important topics of discussion in crisis scenarios [12], event detection and analysis [9], and finding a Twitter user’s home location [5].

Latent Dirichlet Allocation [2], commonly known as LDA, is the dominating topic modeling algorithm. The key advance with LDA is that it models a document as a distribution over topics, and a topic as a distribution over the vocabulary in the corpus. The topic distributions are manually inspected in many studies to show that some underlying pattern exists in the corpus. The meaning behind these topics is often interpreted by the author, and topics are often given a title or name to reflect the author’s understanding of the underlying meaning of the topics. One existing concern with topic modeling algorithms lies with how well human beings can actually understand the topics produced by these algorithms. It may be true that when presented with a group of words a human will always be able to assign some meaning. Measuring the consensus of this meaning against the known underlying properties of the data is the major thrust of this work. We focus on three measures that can be used to calculate the interpretability of topics. First, we reproduce two measures previously introduced in Chang et. al [4]. We find that while these two measures can help to measure the interpretability, they do not show how well the statistical topics illuminate the underlying themes in the data. We continue to introduce Topic Consensus, which measures topic quality based on how well the humans are able to understand the consensus between the topics and the underlying themes. We show that Topic Consensus approximates

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the Owner/Author(s). Copyright is held by the owner/author(s).

HT '15, September 1–4, 2015, Guzelyurt, Northern Cyprus.

ACM 978-1-4503-3395-5/15/09.

<http://dx.doi.org/10.1145/2700171.2791028>.

a different aspect of the LDA topics, helping researchers to better understand the topics generated from their data.

Consensus and interpretability measures for topics rely on crowdsourcing, having human participants answer questions regarding the topics. While this is an important analytical step, these methods can be expensive and time consuming to reproduce, making this analysis out of reach to many researchers. To overcome this problem, we propose a method that uses automatic, computational measures to assess the quality of topics without the need for “Turkers”: the human workers who carry out the tasks for our experiments. We compare the performance of several automated measures introduced previously as well as in this work against those that rely on crowdsourcing to give researchers insight into which measures can be used to assess the quality of their topics. In this way we allow for scalable and cheap methods to assess the topics generated by any topic modeling algorithm.

The main contributions of this work are as follows:

- We propose a new consensus measure, Topic Consensus, that can estimate how well an individual topic represents an underlying topic in the corpus.
- By employing automated measures on the topics, we show that the process of measuring the quality of the topics both from the interpretability and consensus perspectives can be automated. This will allow future researchers to incorporate this topic evaluation framework at scale.

RELATED WORK

Chang et. al conducted the first interpretability study on topic models [4]. In their paper, the authors focus on two main validation schemes for topic models: “Word Intrusion” which studies the top words within a topic by discovering how well participants can identify a word that does not belong. They also introduce “Topic Intrusion”, which studies how well the topic probabilities for a document match with a human’s understanding of this document by showing three highly-probably topics, and one improbably topic, and asking the worker to select the “intruder”. These experiments focus on the humans’ understanding of the dataset. Lau et. al [13] furthers this study by building heuristics to guess the actions of the workers. Other investigations into measuring topic quality include Roder et. al [22], who automatically explore the space of topic quality measures and aggregation functions to find the measure that best approximates model precision. Moreover, Wallach et. al [28] proposes Topic Size and Topic Coherence which we include in this work. Newman et al. [21] proposes a measure for topic coherence based on Pointwise Mutual Information. Furthermore, Aletras and Stevenson [1] measure the quality topics by inspecting the vector similarity. We use these measures by including them in our suite of automated measures that we use to test the crowdsourced measures later in the paper. We assess which of these measures can help us to predict which topics will perform well without the help of the Turkers.

LDA has been used in a wide array of domains. LDA was verified on scientific corpora to show that the topics that were produced by the model made sense [2, 6]. Recently, LDA has been used to study the trends in scientific topics (for example, “Information Retrieval”, and “HIV”) over time, to see how scientific interest has changed over time. LDA has been used in even more varying domains. For example,

Schmidt [23] uses LDA to cluster 1820’s ship voyages. By treating trips as documents and nightly latitude/longitude checkins as words, the authors are able to find trading and whaling topics, amongst others.

Topic analysis can also be used to understand the content of tweets. In the context of disaster-related tweets, Kireyev et. al [12] tries to find disaster-related tweets, finding two types of topics: informational and emotional. Joseph et. al [10] studies the relation between users’ posts and their geolocation. Other works [29, 8] focus on identifying topics in geographical Twitter datasets, looking for topics that pertain to things such as local concerts and periods of mass unrest. In Morstatter et. al [20], the authors use LDA to find evidence of bias in Twitter’s Streaming API.

OVERVIEW OF LATENT DIRICHLET ALLOCATION

LDA discovers “topics” from a large corpus of text. LDA’s definition of a topic is a probability distribution over the corpus vocabulary such that each word in the corpus has some probability of occurring within that topic. LDA will find K topics in the text, where K is a positive integer that indicates the number of topics in the text. K is set by the LDA user, and the process for setting this parameter is largely unknown. LDA will also find the probability of each document to belong to each topic.

LDA takes two inputs: A bag-of-words corpus containing d documents and a vocabulary of size v , and a scalar value K , which indicates the number of topics that the algorithm will learn. LDA then outputs a model, m , which consists of a vector $\mathbf{z}^m$ containing each token’s topic assignment. The output of LDA is twofold:

1. A *Topic* $\times$ *Vocabulary* matrix, $\mathbf{T}^m \in \mathbb{R}^{K \times v}$, with entry $\mathbf{T}_{i,j}^m$ representing the probability of that word j belongs to topic i .
2. A *Document* $\times$ *Topic* matrix, $\mathbf{D}^m \in \mathbb{R}^{d \times K}$, with entry $\mathbf{D}_{i,j}^m$ representing the probability that document i is generated by topic j .

TOPIC INTERPRETABILITY

The key questions evaluated in this work are: 1) are topics produced by LDA interpretable?, and 2) can we automate the process used to evaluate the topics’ interpretability to reduce the dependence on Amazon Mechanical Turk?

To measure the interpretability of topics, we extend upon the framework proposed in Chang et. al [4]. We reproduce the two experiments proposed in their paper: *Word Intrusion*, and *Topic Intrusion*, and continue on to introduce our own measure, *Topic Consensus*. This new measure builds upon the existing crowdsourced methods to tie the topics to a score that measures them based upon their ground truth assignment.

All of the measures used to answer the first question rely on crowdsourcing to verify results. This can be a cumbersome task for researchers, taking ample amounts of both time and money to perform the experiments. In the subsequent section we introduce automatic measures that gauge the quality of topics. We continue to show that not only do these measures approximate the performance of the crowdsourced experiments, but we provide recommendations on

Figure 1: Demographic breakdown of the 188 Turkers who participated in this study.

how to use these automated measures to build better LDA topics for future experiments.

Data

The text corpus focused upon in this study consists of 4,351 abstracts of accepted research proposals to the European Research Council¹. In the first 7 years of its existence, the European Research Council (ERC) has funded approximately 4,500 projects, 4,351 of which are used in this study. Abstracts are limited to 2,000 characters, and when a researcher submits an abstract, they are required to select one of the three scientific domains their research fits into: Life Sciences (LS), Physical Sciences (PE), or Social Sciences and Humanities (SH). These labels will be used in the crowdsourced measure we propose.

Mapping scientific research areas has become of growing interest to scientists, policymakers, funding agencies and industry. Traditional bibliometric data analysis such as co-citation analysis supply us with basic tools to map research fields. However, Social Sciences and Humanities (SH) are proven to be especially difficult to map and to survey, since the fields and disciplines are embedded in diverse and often diverging epistemic cultures. Some are specifically bound to local contexts, languages and terminologies, and the SH domain is lacking coherent referencing bodies or citation indices, and dictionaries [16]. Furthermore, SH terminology is often hard to identify as it resembles everyday speech. Innovative semantic technologies such as topic modeling promise alternative approaches to mapping SH, but the basic question here is: how interpretable are they and how can their results be evaluated in a systematic way. This raises further questions into the interpretability of the LDA topics we study in this paper.

The abstracts used in this research were accepted between 2007 - 2013, written in English, and classified by the authors into one of the three main ERC domains. Table 1 shows some statistics of the corpus. The aim of each abstract is to provide a clear understanding of the objectives and methods to achieve them. Abstracts are also used to find reviewers or match authors to panels.

In addition to the ERC data that lies at the heart of this work, we also created a corpus consisting of the top stories from the New York Times. We do this in the first experiments to act as a control for the specialized topics that may arise from the ERC corpus. These topics are presented to give the reader a point of reference for the ERC topics. The New York Times data was collected using LexisNexis².

¹<http://erc.europa.eu/projects-and-results/erc-funded-projects>

²<http://www.lexisnexis.com/hottopics/lnacademic/>

Table 1: Properties of the European Research Council accepted abstracts and top stories in the New York Times (NYT).

Property	ERC Data	NYT Data
Total Documents	4,351	156
Life Sciences	1,573	—
Physical Sciences	1,964	—
Social Sciences	814	—
Tokens	649,651	9,188
Types	10,016	772

Extracting Topics from Text

We apply LDA to extract topics from the text. We run LDA on the ERC dataset four times, with $K = 10, 25, 50, 100$, yielding a total of 185 ERC topics. Furthermore, we also ran LDA on the New York Times dataset with $K = 25$. All LDA runs were carried out using the Mallet toolkit [17] using the default hyperparameter values of $\alpha = 5.0$ and $\beta = 0.01$.

In addition to the LDA runs, we extract two additional topic groups from the corpus. The first is a set of *random topics*. To generate these topics, we weight the words by their frequency in the corpus and randomly draw words from this distribution. Our topics have a roughly equal mixture of the three ERC domains.

To complement our random topics, we also create a set of *social sciences topics*. To generate these topics, we calculate each word’s probability of occurring in a SH, LS, or PE document. We then select words that occur most in the SH category and least in the other two. These topics lie in contrast to the random topics in that they are strongly skewed to represent a single group.

In both the case of the random and social sciences topics, these topics are different from traditional LDA topics, containing only a *set* of 20 words. For both processes, we generate 25 such topics. We use both of these auxiliary topic sets for validation of the results obtained using the LDA topics. Table 2 shows an overview of all of the topic sets generated for this study.

Amazon Mechanical Turk

The crowdsourced experiments carried out in this work were performed using Amazon’s Mechanical Turk³ platform. Mechanical Turk is a crowdsourcing platform that allows requesters to coordinate Human Intelligence Tasks (HITs) to be solved by Turkers. The formulation of each HIT will be described in the corresponding section for each experiment.

³<http://www.mturk.com>

Table 2: The topic sets used in this study. These indicate the different values of “m” used throughout the experiments.

Name	Dataset	Algorithm	No. Topics
ERC-010	ERC	LDA	10
ERC-025	ERC	LDA	25
ERC-050	ERC	LDA	50
ERC-100	ERC	LDA	100
NYT-025	NYT	LDA	25
RAND-025	ERC	Random	25
SH-011	ERC	Manual	11

In all cases, each HIT was solved 8 times to overcome issues that arise from using non-expert annotators [24].

Prior to solving any HITs, we require the Turker to fill out a demographic survey. The demographic survey consists of five questions about the Turker’s background: their sex, age, first language, country of origin, and highest level of education achieved. In order to understand the demographic makeup of our Turkers, we plot the distribution for each answer in Figure 1. In Figures 1(d) and 1(e), we see a strong skew towards American Turkers who speak English as their first language. This could be partly attributed to a recent change in the Mechanical Turk terms of service that requires Turkers to provide their Social Security Number⁴ in order to solve HITs on the site. This allows us to go forward knowing that the participants are largely English speakers, and we cannot attribute poor performance in our analysis to poor language understanding.

Experiment 1: Word Intrusion

Word Intrusion, introduced by [4], measures the “coherence” of an individual topic. For each topic, we show the Turker the top 5 most probable words from the topic’s probability distribution along with one of the *least* probable words in the distribution (the “intruded” word). We then ask the Turker to select the word that they think is the intruded word. We define model precision as the number of times a Turker was able to guess the intruded word divided by the number of times the HIT was solved, formally:

$$MP_k^m = \frac{1}{|s_k^m|} \sum_{i=1}^{|s_k^m|} \mathbf{1}(g_k^m = s_{k,i}^m), \quad (1)$$

where MP_k^m indicates the model precision k -th topic from model m , g_k^m is the ground-truth intruded word for topic k , and s_k^m is the vector of choices made by Turkers on topic k .

Results

The results of the experiments are shown in Figure 2. We see that the median model precision goes down as a function of K . In [4], we see that the authors attain an average model precision of approximately 90% with both $K = 50$ and 100 on their topics derived from New York Times and Wikipedia data. To compare, our topics achieve median scores of 75% at $K = 50$, and 65% at $K = 100$. These results reflect the difference in Model Precision results that can occur when this methodology is applied to different data.

⁴https://www.mturk.com/mturk/help?helpPage=worker#tax_no_have_tin

Figure 2: Model Precision aggregated by LDA model. Higher is better. We see a trend of decreasing performance with increasing K , and the New York Times corpus performs, on average, worse than the random topics.

Furthermore, see that our reference topics do not perform as expected. While the randomly-generated topics achieve the consistent poor performance that we would expect, the New York Times topics get an even worse score, indicating that humans are worse at detecting an intruded word in our New York Times dataset than they are at random. Furthermore, while the SH topics perform the best of our control topic sets, they still perform worse than any ERC corpus.

Experiment 2: Topic Intrusion

Topic intrusion, introduced by Chang et. al [4], measures how well the *Document* $\times$ *Topic* matrix, $\mathbf{D}^m$, produced by the LDA model can be deciphered by humans. This matrix is a reflection of how well each document is described by each topic produced by the model. Each row of this matrix $\mathbf{D}_d^m$ is a probability distribution over all of the topics with entry $\mathbf{D}_{d,i}^m$ being the probability that document d is made up of topic i .

In this task, a document is shown to the Turker, along with three highly-probable topics and one improbable topic. The highly-probable topics are selected as the three highest values in $\mathbf{D}_d^m$, and the improbable topic is randomly select from the bottom three non-zero values of $\mathbf{D}_d^m$. The Turker is told to identify the improbable, or intruded, topic. To calculate the error, we adopt the Topic Log Odds (TLO) calculation, defined as follows:

$$TLO_d^m = \frac{1}{|s_d^m|} \sum_{i=1}^{|s_d^m|} (\log(\mathbf{D}_{d,s^*}^m) - \log(\mathbf{D}_{d,s_d^m,i}^m)), \quad (2)$$

Figure 3: Topic Log Odds on the ERC and New York Times data. Higher is better.

where $\mathbf{s}_d^m$ is the vector of Turker answers for document d on model m , s^* is the index of the true intruded topic, and $\mathbf{s}_d^m, i$ is the index of the i -th Turker’s answer.

Results

The results of the experiments are shown in Figure 3. Due to the way the SH and Random topics were generated, we were unable to perform this analysis on these datasets. In Chang et. al [4], the Turkers get median scores of -1.5, and -1.6 for $K = 50$, and $K = 100$, respectively. To compare we get results of -2.1, and -1.8 for these groups. Again, we compare with [4] to show how a different dataset can influence the results of these measures.

Experiment 3: Topic Consensus

The previous experiments test the intrinsic coherence of individual topics using Word Intrusion, and the fit of the topics to the documents using Topic Intrusion. We have not yet measured the ability of the topics to conform to the natural topics underlying the text. Explicit topics are often present in many corpora, such as newspaper articles, due to their manual categorization by “sections”, and our ERC abstracts which explicitly label each abstract with an ERC domain. We measure this conformity to the underlying topic distribution by leveraging the ground-truth topic labels that comes from the ERC domains when the abstract is submitted.

To understand how well the statistical topics mimic the underlying topics, we show the Turker the top 25 words of a statistical topic and ask him to choose which of the three ERC domains the topic describes: “Life Sciences”, “Physical Sciences”, or “Social Sciences”. We provide a fourth option, “No Topic Matched”, in case the topic does not make sense to the Turker.

To compute the topic’s ability to represent the underlying topics in the dataset, we compare the distribution of the Turkers’ responses for that topic with the distribution of the topic over the ERC domains. To perform this analysis, we construct an LDA Topic $\times$ ERC domain matrix $\mathbf{R}$, where $\mathbf{R}_{i,c}$ indicates topic i ’s probability of occurring in ERC domain c . By defining $\mathbf{R}$ in this way, we are able to obtain values for topics that were not generated through LDA. The structure of each row of $\mathbf{R}$ is dependent on the type of topic group it comes from. We construct the rows as follows for each topic group:

- **ERC-*** — The $\mathbf{R}_i$ row vector for an ERC topic is created by taking the sum of the columns of the $\mathbf{D}$ matrix. This sum is taken for each row (document) of $\mathbf{D}$ labeled with the corresponding ERC domain. This is defined as:

$$\mathbf{R}_{i,c} = \frac{\sum_{j \in M_c} \mathbf{D}_{j,i}}{\sum \mathbf{D}_{*,i}}, \quad (3)$$

where M_c is the set of documents containing the label corresponding to the column of $\mathbf{R}$, i.e. “SH”, “LS”, or “PE”. This gives us an understanding of the ERC domain makeup for each LDA topic.

- **SH-011** — The $\mathbf{R}_i$ row vector for an SH topic contains a 1 for the SH category and a 0 for the other two. This is because due to the way the SH topics are generated, they contain purely SH words.
- **NYT-025** and **RAND-025** — Both Random and NYT topics have an evenly split row vector of 1/3 for each category as they do not belong to any category. The aggregate of the Turkers’ responses should be confused over all of the ERC categories. We will measure this disparity in subsequent sections.

Using the responses from the Turkers, we build a separate Topic $\times$ Domain matrix, $\mathbf{R}^{\text{AMT}}$ where $\mathbf{R}^{\text{AMT}}_{i,j}$ represents the Turkers’ probability of choosing domain j when presented with topic i . In this way, $\mathbf{R}^{\text{AMT}}$ is the representation of $\mathbf{R}$ obtained from the Turkers’ responses. A row in $\mathbf{R}^{\text{AMT}}$ indicates the distribution over ERC domains for a given LDA topic from the Turker’s responses.

To compute the topic consensus of the topics, we compute the Jensen-Shannon divergence [14] between the two distributions $JS(\mathbf{R}^{\text{AMT}}_d || \mathbf{R}_d)$, defined as:

$$JS(\mathbf{R}^{\text{AMT}}_d || \mathbf{R}_d) = \frac{K(\mathbf{R}^{\text{AMT}}_d || M) + K(\mathbf{R}_d || M)}{2}, \quad (4)$$

where K is Kullback-Leibler divergence [11], and $M = \frac{1}{2}(\mathbf{R}^{\text{AMT}}_d + \mathbf{R}_d)$. Jensen-Shannon is a natural choice as the rows of $\mathbf{R}$ and $\mathbf{R}^{\text{AMT}}$ are probability distributions over the 3 ERC topics and Jensen-Shannon is a measure of the similarity of two distributions.

Results

The results of the topic consensus experiment are shown in Figure 4, and a confusion matrix comparing the Turkers’ responses with the ground truth is shown in Table 3. In Figure 4, we see that the SH topics perform the best, showing that Turkers are able to match these handpicked words with the topic they come from. The New York Times topics perform worst of all. This is partly explained by Table 3,

Table 3: Confusion matrix of ground truth ERC domain assignments of topics against the domain assignments made by the Turkers. In the ERC topics, we see that the Turkers are generally able to identify SH and LS topics, but overall fail to identify PE topics. The Turkers perform well when shown random topics, giving most of these topics a “not applicable” label.

	ERC Topics				NYT Topics			
	Ground Truth				Ground Truth			
	LS	SH	PE	NA	LS	SH	PE	NA
AMT Classification								
LS	45	0	2	0	0	0	0	1
SH	3	10	8	0	0	0	0	22
PE	1	0	60	0	0	0	0	0
NA	4	0	52	0	0	0	0	2
Accuracy	62%				8%			
	Random Topics				SH Topics			
	LS	SH	PE	NA	LS	SH	PE	NA
	LS	SH	PE	NA	LS	SH	PE	NA
LS	0	0	0	1	0	0	0	0
SH	0	0	0	5	0	11	0	0
PE	0	0	0	0	0	0	0	0
NA	0	0	0	19	0	0	0	0
Accuracy	76%				100%			

where we see that many NYT topics are classified as SH by the Turker, indicating that the Turkers may have an over-interpretation of what constitutes an SH topic. This could also help to explain their poor performance in the Word Intrusion task.

The Difference between Topic Measures

We have executed three measures on our data thus far: Word Intrusion, Topic Intrusion, and Topic Consensus. At this point we should step back and discuss the differences observed between these three measures. First, in Word Intrusion, we measured the Turker’s ability to identify a word randomly inserted into the top words in a topic. This measure approximates a topic’s intelligibility to the reader. Next, we studied Topic Intrusion, in which we saw how well the user could understand the Document by Topic matrix $\mathbf{D}$ generated by LDA. Both of these measures were proposed in Chang et. al [4].

After reproducing these two measures, we continued to propose our own Topic Consensus measure to assess the quality of the topics. The main difference between this measure is that it refers to properties of the corpus from which the topics were generated in order to compare the quality of the topics. Simultaneously, it helps us to identify topics whose makeup of real, corpus-level groups are discernible to the reader. This measure can be useful for statistical topics generated on any corpus with corpus-level groups, such as newspaper articles or even Wikipedia entries.

In our crowdsourced experiments, we see that the Turkers perform admirably. We also see that the Turkers performance differs from that obtained in [4]. This is likely due to the nature of the data used in our study. Clearly, there is a need for automated measures that can supplement the knowledge of the Turkers. In the next section we test a suite of possible measures for measuring topic quality.

MEASURING TOPIC QUALITY AUTOMATICALLY

In the previous sections we reproduced and introduced a variety of approaches to measure the coherence and consensus of topics. All of these measures rely on crowdsourcing to get the results, which can cause for a bottleneck in the research process. Automating the task of the Turkers can make for more accessible empirical comparison, saving researchers both the time and cost of performing crowdsourced experiments. In this section we compare measures that can be used to automatically discover topics that are of high quality according to the measures above.

Automatic Topic Quality Measures

Here, we discuss several methods for automatically assessing the quality of topics. Each measure is defined as follows:

1. **Topic Size (TS):** This is a count of the number of tokens in the input corpus that are assigned to the topic. This was used in Mimno et. al [18] as a possible measure for topic quality. The hypothesis behind this measure is that a larger topic (with more tokens) will represent more of the corpus, and thus convey a larger understanding.
2. **Topic Coherence (TC):** Also introduced by Mimno et. al [18], this measures the probability of top words co-occurring within documents in the corpus:

$$TC(\mathbf{w}) = \sum_{j=2}^{|\mathbf{w}|} \sum_{k=1}^{j-1} \log \frac{D(\mathbf{w}_j, \mathbf{w}_k) + 1}{D(\mathbf{w}_k)}, \quad (5)$$

where $\mathbf{w}$ is a vector of the top words in the topic sorted in descending order, and D is the number of documents containing all of the words provided as arguments. This measure is computed on the top 20 words of the topic.

Figure 4: Topic consensus scores across all models. Lower scores are better. On the left we see the form that is shown to the workers. On the right we see that the random topics perform worse than any of the ERC topics, and the SH topics perform the best.

3. **Topic Coherence Significance (TCS):** We adapt the measure above to understand the significance of the top words in the topic when compared to a random set of words. To calculate this measure we select 100 groups of words at random, following the topic’s word distribution. We then recompute the Topic Coherence measure for each of the random topics, obtaining a vector, $\mathbf{d}$, of topic coherence scores. We calculate the mean, $\bar{\mathbf{d}}$, and standard deviation, $std(\mathbf{d})$. Significance is defined as:

$$TCS(\mathbf{w}) = \frac{TC(\mathbf{w}) - \bar{\mathbf{d}}}{std(\mathbf{d})}. \quad (6)$$

4. **Normalized Pointwise Mutual Information (NPMI):** Introduced by Bouma [3], this metric measures the probability that two random variables coincide. This measure was used to estimate the performance of Model Precision in Lau et. al [13], where the authors adapted it to measure the coincidence of the top $|\mathbf{w}|$ words. In this paper, two variables “coinciding” is the probability that they will co-occur in a document. The authors named this version OC-Auto-NPMI, formally:

$$OC\text{-}Auto\text{-}NPMI(\mathbf{w}) = \sum_{j=2}^{|\mathbf{w}|} \sum_{k=1}^{j-1} \frac{\log \frac{P_D(\mathbf{w}_j, \mathbf{w}_k)}{P_D(\mathbf{w}_j)P_D(\mathbf{w}_k)}}{-\log(P_D(\mathbf{w}_j, \mathbf{w}_k))}, \quad (7)$$

where $P_D(\cdot) = D(\cdot)/|N|$, where $|N|$ is the number of documents in the corpus. P_D measures the probability that a document in the corpus contains the words given to $D(\cdot)$. We will refer to this measure as NPMI.

5. **ERC-Distribution HHI (ERC-HHI):** The Herfindahl-Hirschman Index [7], or HHI, is a measure proposed to find the amount of competition in a market. This is calculated by measuring the market share of each firm in the market, formally:

$$HHI = \frac{(H - 1/N)}{(1 - 1/N)}, \quad (8)$$

Where $H = \sum_i s_i^2$, N is the number of firms in the market, and s_i is the market share of firm i , as a percentage. HHI ranges from 0 to 1, with 1 being a perfect monopoly (no competition), and 0 being an evenly split market. In this way we measure how focused the market is on a particular firm.

By treating ERC categories as firms, and the market share distribution as $\mathbf{R}_i$, we can calculate how focused each topic is around a particular ERC domain.

6. **Topic-Probability HHI (TP-HHI):** Using the same formulation as ERC-HHI, this time we treat every *word* in the vocabulary as a firm, and $\mathbf{T}_i^m$ as the probability distribution. In other words, the topic’s probability distribution is the market, and TP-HHI measures the market’s focus any word, or group of words. This measures whether the focus of a topic is around a handful of words, or whether it is evenly spread across the entire vocabulary used to train the model.

Correlation with Crowdsourced Measures

To see how well these automatic topic measures compare with the crowdsourced topic measures from the “Topic Interpretability” section, we calculate the Spearman’s ρ [25] rank correlation coefficient between the crowdsourced measure and the automatic measure. We omit the “topic log odds” score as it is not a true topic measure, but instead a measure of the documents over topics. The correlations between each pair of crowdsourced and automatic measure are shown in Table 4. In this table, we present the Spearman’s ρ , as well an indication of the significance for the following hypothesis test:

H_0 : The two sets of data are uncorrelated.

Instances where the hypothesis is rejected at the $\alpha = 0.05$ significance level are shown in the table.

In Table 4, we see that the measure most correlated with Model Precision is NPMI, meaning that documents whose words co-occur more are apt to achieve higher Model Precision values. We see that higher (better) values of model precision are accompanied by lower NPMI values.

On the other hand, we see that the measure most correlated with Topic Consensus is TP-HHI. Once again, these lists exhibit negative correlation, indicating that the more peaked a distribution is, the lower (better) the Topic Consensus will be. This is a hopeful result for researchers, as the measure does not rely on the topic’s domain distribution information, meaning that an approximate ranking of the topics could be obtained *without* ground truth.

Precision with Crowdsourced Measures

In the previous section correlation was used to determine the dependence between the automated measures and the crowdsourced measures. To accompany this analysis, we also want to see the automated measures’ *precision* in predicting the best topics from the LDA model. To measure this, we

Figure 5: Agreement between the automated measures and crowdsourced measures for different values of k . In the case of Topic Coherence, TP-HHI yields the best agreement. In the case of Model Precision, the Topic Coherence does best at the shorter lists, however is matched by other measures at longer lists.

Table 4: Spearman’s ρ between measures requiring Mechanical Turk and automated measures. Instances where we reject the null hypothesis at the $\alpha = 0.05$ significance level are denoted with *, and instances where we reject the null hypothesis at the $\alpha = 10^{-6}$ significance level are denoted with **.

	Model Precision	Topic Consensus
TS	0.152*	-0.532**
TC	0.359**	-0.584**
TCS	-0.074	-0.788**
ERC-HHI	0.103	-0.478**
TP-HHI	-0.471**	-0.885**
NPMI	-0.562**	-0.774**

calculate the agreement between the top k topics determined by ranking the crowdsourced measures as well as the top k topics determined by ranking the automated measures. The results of this experiment are shown in Figure 5. Here we see that **TP-HHI** is the best at finding the best topics at lists of any size with respect to Topic Coherence. In the case of model precision the best measure is less clear: **Topic Coherence** is best for shorter lists while **TP-HHI**, and **Topic Size** compete for the best performance of some of the mid-length lists.

Predicting the True Crowdsourced Values

So far we have evaluated the quality of our automated measures by seeing how well the value of the automated measures changes as compared to the crowdsourced measures. While correlation can be used to find the quality of the measure, and precision to find the best topics, there is value in being able to predict the *actual* value of the topics’ crowdsourced measures.

Using linear regression, we train a model to predict the true value of the crowdsourced measures using the automated measures. We build two models: one where the dependent variable is “Model Precision”, and another where the dependent variable is “Topic Coherence”. In both cases, the independent variables are all of the automated measures introduced previously. We use 10-fold cross validation and report both the mean and standard deviation of the performance of the models across all 10 runs. In this experiment we achieved a root-mean-square error of 0.12 ± 0.02 for Topic Coherence and 0.27 ± 0.05 for Model Precision.

CONCLUSION AND FUTURE WORK

We investigate how well topics can be interpreted by humans. To do this we generated topics from a collection of scientific abstracts, consisting of abstracts of proposals accepted by the European Research Council. Using this data we ask how well crowdsourced techniques can be used to evaluate topic models. Further, we investigate how well automatic measures can approximate the crowdsourced measures, making way for scalable and reproducible research.

We reproduced existing topic measures on a corpus consisting of scientific abstracts. We received inferior results compared to the original results which were performed on Wikipedia and New York Times data. This could be due to the nature of the Amazon Mechanical Turk workers changing, but is likely due to the nature of the dataset.

We propose a new crowdsourced topic measure, “topic consensus”, that includes ground truth information about the label of the topic. This method allows researchers to see how well the topics generated from their dataset can be understood in relation to the underlying topics in the corpus. Furthermore, by inspecting the Turker’s results on the NYT topics we were able to see how the Turker’s understanding of the topics influences their answers.

We investigate how to estimate these crowdsourced topic measures without the need of crowdsourcing tools such as Mechanical Turk. We find some automated measures that

are very highly correlated with crowdsourced measures, allowing future researchers to reproduce these topic quality measures at scale. Additionally, we find that the measures that best estimate the ground truth topics do not require ground truth information themselves, allowing for researchers to perform topic consensus analysis when ground truth information about the underlying ground truth topics in their dataset is not known.

One direction for future work is to catalogue the performance of all crowdsourced measures across different datasets to provide a benchmark for different types of data. Future work also includes seeing how these measures perform when K is taken out of the equation, using nonparametric topic modeling algorithms such as Hierarchical Dirichlet Processes [26]. Other areas of interest include the role that both “interpretability”, and “coherence” contribute to the measures proposed in this work.

ACKNOWLEDGEMENTS

This work is sponsored, in part, by Office of Naval Research grant N000141410095 as well as LexisNexis and HPCC Systems.

REFERENCES

- [1] N. Aletras and M. Stevenson. Evaluating Topic Coherence using Distributional Semantics. In *IWCS*, pages 13–22, 2013.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [3] G. Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, pages 31–40, 2009.
- [4] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*, volume 22, pages 288–296, 2009.
- [5] J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *EMNLP*, pages 1277–1287, 2010.
- [6] T. L. Griffiths and M. Steyvers. Finding Scientific Topics. *PNAS*, 101(Suppl 1):5228–5235, 2004.
- [7] A. O. Hirschman. *National Power and the Structure of Foreign Trade*. University of California Press, Berkeley, CA, 1945.
- [8] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsoulis. Discovering Geographical Topics in the Twitter Stream. In *WWW*, pages 769–778, 2012.
- [9] Y. Hu, A. John, F. Wang, and S. Kambhampati. ET-LDA: Joint Topic Modeling for Aligning Events and their Twitter Feedback. In *AAAI*, volume 12, pages 59–65, 2012.
- [10] K. Joseph, C. H. Tan, and K. M. Carley. Beyond “Local”, “Categories” and “Friends”: Clustering foursquare Users with Latent “Topics”. In *UbiComp*, pages 919–926, 2012.
- [11] J. M. Joyce. Kullback-Leibler Divergence. In *International Encyclopedia of Statistical Science*, pages 720–722. Springer, 2011.
- [12] K. Kireyev, L. Palen, and K. Anderson. Applications of Topics Models to Analysis of Disaster-Related Twitter Data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, volume 1, 2009.
- [13] J. H. Lau, D. Newman, and T. Baldwin. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.
- [14] J. Lin. Divergence Measures Based on the Shannon Entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, Jan 1991.
- [15] J. Mahmud, J. Nichols, and C. Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users. In *ICWSM*, pages 511–514, 2012.
- [16] K. Mayer and J. Pfeffer. Mapping Social Sciences and Humanities. *Horizons for Social Sciences and Humanities*, 09 2014.
- [17] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>, 2002.
- [18] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing Semantic Coherence in Topic Models. In *EMNLP*, pages 262–272, 2011.
- [19] F. Morstatter, N. Lubold, H. Pon-Barry, J. Pfeffer, and H. Liu. Finding Eyewitness Tweets During Crises. In *ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014.
- [20] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. In *Proceedings of ICWSM*, pages 400–408, 2013.
- [21] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic Evaluation of Topic Coherence. In *NAACL*, pages 100–108, 2010.
- [22] M. Röder, A. Both, and A. Hinneburg. Exploring the Space of Topic Coherence Measures. In *WSDM*, pages 399–408, 2015.
- [23] B. M. Schmidt. Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, 2(1):49–65, 2012.
- [24] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *EMNLP*, pages 254–263. Association for Computational Linguistics, 2008.
- [25] C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904.
- [26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), 2006.
- [27] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welp. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*, pages 178–185, 2010.
- [28] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation Methods for Topic Models. In *ICML*, pages 1105–1112, 2009.
- [29] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical Topic Discovery and Comparison. In *WWW*, pages 247–256, 2011.