A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability

AMIN HOSSEINY MARANI[†], Lehigh University, CSE Department, U.S
ERIC P. S. BAUMER [†], Lehigh University, CSE Department, U.S

Topic modeling includes a variety of machine learning techniques for identifying latent themes in a corpus of documents. Generating an exact solution (i.e., finding global optimum) is often computationally intractable. Various optimization techniques (e.g., Variational Bayes or Gibbs Sampling) are employed to generate topic solutions approximately by finding local optima. Such an approximation often begins with a random initialization, which leads to different results with different initializations. The term "stability" refers to a topic model's ability to produce solutions that are partially or completely identical across multiple runs with different random initializations. Although a variety of work has been done analyzing, measuring, or improving stability, no single paper has provided a thorough review of different stability metrics nor of various techniques that improved the stability of a topic model. This paper fills that gap and provides a systematic review of different approaches to measure stability and of various techniques that are intended to improve stability. It also describes differences and similarities between stability measures and other metrics (e.g., generality, coherence). Finally, the paper discusses the importance of analyzing both stability and quality metrics to assess and to compare topic models.

CCS Concepts: • **Computing methodologies** → **Topic modeling**; • **General and reference** → **Evaluation**.

Additional Key Words and Phrases: topic modeling stability

1 INTRODUCTION

Topic models are a series of unsupervised machine learning techniques that process text datasets with large number of documents to extract latent themes [12, 100]. Although technically these techniques can be applied to various types of data (e.g., text vectors, image pixels, or even potsherds [23, 83]), topic modeling is often employed to extract topics from a large corpus of text documents [47, 66, 100, 125]. Generally, a topic is represented as a probability distribution over a vocabulary of words. Each document is then represented as a probability distribution over topics, i.e., each document is a probabilistic mixture of topics. Any topic modeling technique generates these distributions based on observed document-term information (and in some models with metadata) [12, 100, 121]. Each of these different techniques infers the unobserved distributions (i.e., topic-term probability and topic-document distribution) in slightly different ways.

However, running a single model multiple times with the same input documents will generate different topic solutions. This issue, referred to as *stability*, occurs with many machine learning techniques that are initialized randomly or take random steps to generate solutions [16, 34, 115]. While there were attempts to solve instability

[†]This material is based upon work supported by the National Science Foundation under NSF Grant #1816403.

Authors' addresses: Amin Hosseiny Marani[†], Lehigh University, CSE Department, Building C, 322 Research Drive, Bethlehem, U.S, amh418@lehig.edu; Eric P. S. Baumer [†], Lehigh University, CSE Department, Building C, 113 Research Drive, Bethlehem, U.S, ericpsb@lehigh.edu.

0360-0300/2023/9-ART $15.00
https://doi.org/10.1145/3623269

in machine learning models (e.g., KNN [48], Bagging [63], Random Forests [37], Stochastic Gradient Descent [43], K-means [57], and Neural Networks [58]), there is no single solution that can be used for all models [49, 103].

Topic modeling has a wide range of applications in NLP, including information retrieval [117], summarization [8], text classification [19, 90], machine translation [95], and others. Across these applications, stability poses unique challenges. A topic model that lacks stability may result in uncertainty in generating outputs (e.g., using topics as features for classification or obtaining themes of a corpus). Furthermore, topic modeling has also been applied in other domains, including historical document analysis [42, 122, 124], humanistic analysis of literature [65, 68], qualitative and quantitative social science [10], recommender systems [4, 122], interactive visualizations [8, 108], and many other purposes [18, 77]. Without stability, numerous issues emerge: visualizations become inconsistent, social scientific analyses may lack replicability, and interpretations of what a topic is about can become uncertain or confusing [9, 26, 98].

Despite introducing different stability measures, no prior work has organized these measures and their properties. A systematic review of these measures and their properties is both valuable and important, for at least two main reasons. First, choosing a stability measure from dozens of approaches poses a complex and challenging task, as it requires understanding what each measure offers, what are unique properties of each measure, and what input each measure needs. Thus, a review of stability measures will support researchers to compare stability measures regarding these different aspects.

Second, reviewing prior work designed to improve topic modeling techniques regarding their stability will also help to support researchers to employ and compare models with a higher stability suited to their tasks and applications. For example, running topic modeling techniques multiple times for interactive visualization applications or quantitative social science analysis will result in slightly different results that needs a new interpretation or investigation. Reviewing prior work that improved topic modeling techniques regarding their stability will help to support researchers in order to employ and compare models with a higher stability suited to their tasks and applications.

This paper organizes prior work on addressing stability in topic modeling around two general areas. First, *measuring stability* provides a quantitative metric to determine how stable the results are across runs. Generally, stability is measured by capturing changes in topic solutions across multiple runs with different random initializations of a given model on a given data set [28, 70, 112]. After briefly describing dominant topic models (Section 2) and summarizing prior review papers on topic modeling (Section 3), Section 5 surveys a variety of different metrics for computing the stability of topic modeling results.

Second, once stability metrics are selected, various techniques can be explored for *improving stability*. Examples include using different inference [100] or sampling processes [53], aggregating across multiple random initializations [56], proposing entirely different models [32], and others. Section 6 surveys a variety of these techniques, comparing and contrasting not only the techniques themselves but also their impact on stability.

While improving stability can provide a number of benefits, as described above, it is also important to examine relationships between stability and other properties of topic modeling. For example, prior work has explored various approaches to assessing the quality of topic modeling results [22, 61, 75, 121]. Just as a lack of stability leads to different results across different random initializations, these quality metrics may also vary across initializations [50, 51]. Furthermore, stability itself could be used as a quality metric, perhaps in part to select the number of topics [26, 32, 41]. Section 7 discusses such relationships in more detail, as well as potential trade-offs between improving stability and other aspects of topic modeling. This discussion thus lays important groundwork for future research on topic modeling stability.

This paper reviews the sources of instability in topic modeling techniques in Section 2. It also addresses the gap in prior work regarding to reviewing stability in Section 3. The method section (i.e., Section 4) discusses the review process, including identification, screening, and inclusion-exclusion processes. To reiterate, this paper's contributions include: a survey of metrics used to measure stability (Section 5), a review of how effectively

different techniques improve stability (Section 6), and suggestions for future work to examine the relationships among stability and other properties of topic models (Section 7).

2 TOPIC MODELING

This section reviews four popular and early topic models: Latent Semantic Analysis (LSA), probabilistic LSA (pLSA), Latent Dirichlet Allocation (LDA), and Non-negative Matrix Factorization (NMF) to describe how topic modeling works in general. There exist more than a dozen topic modeling techniques, and more than a hundred topic modeling extensions, thus it is difficult and perhaps impossible to cover them all in one section. Furthermore, most models are not designed to resolve topic instability, which is the main focus of this paper. Later, Section 5 discusses what causes instability, and how different approaches tackle this problem based on these descriptions.

Deerwester et al. [29] introduced a model, called Latent Semantic Analysis (LSA), to extract k-dimension term vectors from documents of a corpus. LSA first builds a document-term frequency matrix X of size $t * d$, where t is the number of terms in the vocabulary and d is the number of documents in the corpus. Term frequencies can be either raw counts or some transformation, e.g., log frequency [30], log entropy, or tf-idf [84]. A Singular Value Decomposition (SVD) is then applied on X, which generates three matrices: T (a term matrix of size $t * r$), S (a diagonal matrix of size $r * r$), and D (a document matrix of size $d * r$) in which r is the number of singular values. These resultant matrices can be used to reconstruct the term-document frequency matrix X by multiplication with minimum error. LSA keeps the K largest singular values of matrix S, as well as their associated K dimensions in the other vectors. Thus, the first K columns of matrix T represent a term-topic matrix, and the first K rows of matrix D represent the topic-document features. Due to the stochastic nature of SVD [106], applying LSA to the same corpus multiple times, including running SVD and retaining only the dimensions with the largest singular values, will result in different topic solutions, i.e., a lack of stability. Moreover, SVD generates both negative and positive values, but interpreting negative probabilities is complicated [111].

pLSA is a probabilistic version of Latent Semantic Analysis (LSA). pLSA resolves the negative values problem by building a generative topic model of documents. The main assumption of pLSA is that each document is sampled from the document-topic distribution and is generated by randomly sampling words using topic-term probabilities. Therefore, pLSA assigns words and documents to topics by maximizing the log-likelihood of the document-term matrix. Figure 1 shows the graphical model of the pLSA, in which node w represents observable words of documents, node z indicates latent topics, and node d indicates documents for which the topic proportions are not observable. In this figure, $P(z|d)$ indicates the probability of a topic z appearing in a given document d, and $P(w|z)$ shows the probability of a word w to be drawn randomly given a topic z.

Later, Blei et al. [15] added a hyperparameter α, a prior Dirichlet Allocation, to the pLSA topic model. They also used Gibbs sampling [39, 92] to optimize topic generation in order to improve both term-topic and document-topic assignment regarding a defined fitness function. This method is called Latent Dirichlet Allocation (LDA). LDA's generative process for a document starts with choosing document length from a Poisson distribution. Then, the topic proportion for each document is drawn from a Dirichlet distribution ($\theta = Dir(\alpha)$). Finally, for each document, a topic is chosen with multinomial distribution over previously obtained θ, and words are drawn from topic-term probability matrix β for that document [15]. Figure 2 shows the graphical representation of the LDA topic model. Using randomized learning algorithms (e.g., Gibbs sampling [92], variational Bayes [36], or Expectation Maximization [110]) in the topic generation process, random initialization of β and θ, and possibly changing hyperparameter α, yield different topic solutions.

As an alternative to LDA, Non-negative Matrix Factorization (NMF) is a topic modeling approach that resolves the above-mentioned negative values generated from decomposed matrices. Paatero and Tapper [86] first introduced a matrix factorization algorithm with non-negative constraints called positive matrix factorization to avoid generating negative values. Arora et al. [7] used non-negative SVD and introduced NMF topic modeling.

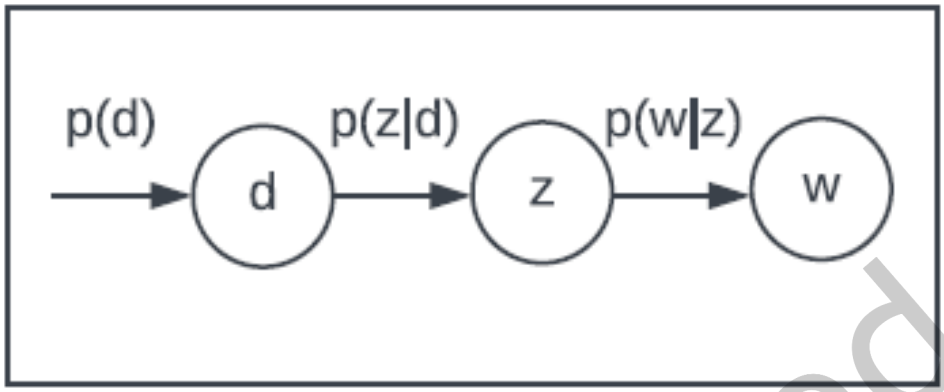

Fig. 1. The graphical representation of probabilistic Latent Semantic Analysis. Each node is a random variable with an assigned label that shows its role in the topic generation procedure [44].

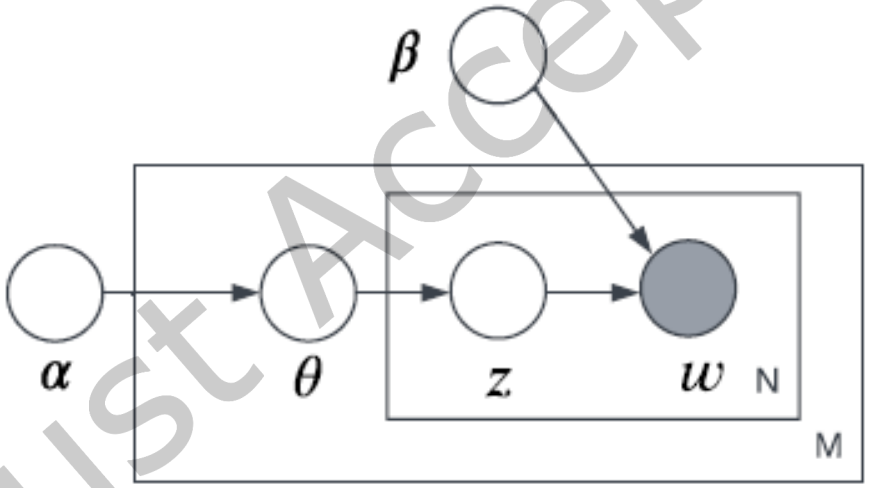

Fig. 2. The graphical model of LDA. Node w is the only observed variable which is shaded. N plate represents word-topic space and M plate represents document space of LDA [15].

Figure 3 shows a schematic of NMF topic modeling. Unlike LSA, NMF does not decompose the document-term matrix into r dimension and take K dimensions with largest singular values. It decomposes into K dimensions as K topics.

Following the development of these initial topic models, a series of new methods were introduced to make various improvements (e.g., improving topic quality, topic stability, or convergence speed). These methods can be grouped into two categories; matrix factorization and generative models. Matrix factorization methods do not

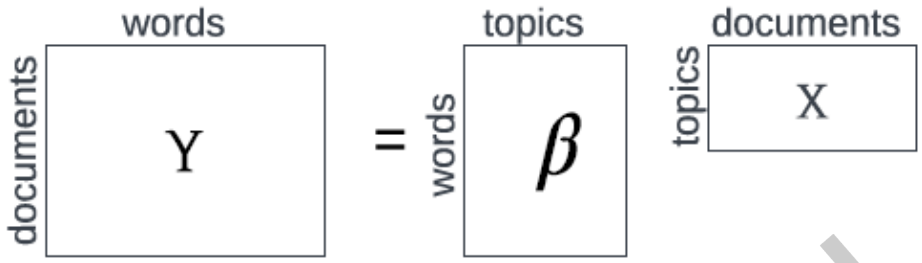

Fig. 3. Matrix factorization topic modeling schema [28]. Matrix factorization techniques obtain latent topic-document and topic-term matrices using the observed information, document-term matrix.

take priors (such as α in LDA) and do not generate topics based on a probability model or Bayesian inference. These method decompose a document-word matrix with different objective and optimization methods to generate topic-word and document-topic matrices. On the other hand, generative models produce topics to minimize the error between the original documents and documents generated by the topic generative process [12]. These topic modeling techniques are based on a probability model [13, 73, 92].

Since topic models extract unobserved latent topics using randomized algorithms, regardless of the topic model approach, running a model multiple times result in different solutions. In LDA or PLSA random initialization has an impact on the final topic solutions. Besides random initialization, in Matrix Factorization approaches (e.g., LSA, or non-negative SVD) different orders of documents in document-term matrix influence the decomposition process and extracted topics [5]. Measuring these differences (i.e., stability), finding sources of these variations, and fixing this problem are the focus of the sections 5 and 6. Next section reviews the previous surveys and review papers of topic modeling methods to point out an existing gap in reviewing stability issue in topic modeling.

3 RELATED WORK: TOPIC MODELING SURVEYS

Over half a dozen papers have reviewed different topic modeling approaches through varied lenses, from topic modeling techniques [27, 47, 107], to topic modeling applications [3, 18, 94, 101, 125], to properties and diagnostics [125]. Each of these sought to summarize a variety of topic models, comparing and contrasting different approaches, applications, or capabilities.

One of the earliest surveys was provided by Daud et al. [27] on the topic modeling methods. Their paper briefly describes the topic modeling process from text preprocessing to representing topics as top terms. It introduces topic modeling, reviews the optimization procedure in general, covers different inference algorithms, summarizes some applications of topic modeling. Later, Sharma et al. [107] provided a more comprehensive survey of topic models, including LDA, LSA, PLSA, and matrix factorization models (e.g., SVD). They also cover n-gram variant models (e.g., Bigram-LDA [80]), deep learning topic models (e.g., Deep Belief Networks [67]), and non-LDA topic models (e.g., Aspect Hidden Markov Model [14]) that use different word prepossessing, inference, and topic generation algorithms.

Another group of survey papers focused instead on reviewing topic modeling applications. Boyd-Graber et al. [18] gathered a comprehensive survey of topic modeling applications, focusing on LDA and its extensions. Examples range from summarizing scientific, historical, and fictional documents; to social scientific uses; to multilingual and machine translation applications. Similarly, Wesslen [125] reviewed different social scientific applications of

topic modeling and contrasted multiple topic models, including LDA, Correlated Topic Modeling [13], Dirichlet Multinomial Model [74], and Structural Topic Modeling [100].

Topic models have been used to explore documents of a corpus with metadata about time (e.g., publication dates for books, timestamps for social media posts). While traditional topic modeling techniques (e.g., LDA, PLSA, and LSA) do not take time/date meta-data into inferencing phase, researchers developed dynamic time topic models to fill this gap (e.g., Structural Topic Modeling [100]). Alghamdi and Alfalqi [3] reviewed topic evolution models that considers timeline in inference process and contrast abilities of these models to traditional topic models.

Given the large volume of papers, new techniques, and extensions of topic modeling, many reviews focus on properties and diagnostics of topic modeling methods. Jelodar et al. [47] surveyed recent LDA topic models, and summarized inference algorithms of LDA models (e.g., Variational Bayes, Gibbs sampling, and Expectation Maximization) within different timelines (e.g., research between 2010 to 2011, or 2013 to 2014). A large proportion of this survey outlined different applications of LDA models including linguistic science, medical and biomedical, political science, geographical and locations, software engineering, social media, and crime prediction applications.

With various topic models, performance measurement becomes important to help researchers select one model over another, or to tune a model's (hyper)parameters. A wide variety of topic modeling's performance assessment has been introduced from mathematical perspective of how a model is a good fit on the unseen data (e.g., perplexity or held-out likelihood [120]), to mutual information evaluation techniques (e.g., NPMI or PMI [61]), to single-item likert scale human assessment [61, 79]. Röder et al. [101] reviewed a handful of Coherence scoring schemes (e.g., PMI, NPMI, UCI, and UMASS [111]). Coherence scoring schemes that are based on mutual information of top-term pairs have been receiving more attention as these metrics show strong correlation with human assessment [61]. Their paper analyzes different metrics' performance across varied reference corpora and compare scoring methods' correlation to human assessment.

Stability issues, both with machine learning generally and topic modeling more specifically, have been described, analyzed, and discussed in many papers [11, 28, 34, 53, 69, 97]. While there are a handful of survey and review papers about topic models, a comprehensive study of stability's importance, metrics for measuring stability, and techniques for improving stability has not been published yet. This paper fills that gap. It reviews sources of instability in topic models, compares stability measures and similarity metrics, and discusses stability-related improvements in topic modeling. It concludes by contrasting stability measures to other topic modeling evaluation measures.

4 METHODS

To reiterate, the goal of this review paper is to provide reviewers, authors, and researchers an understanding of different approaches to measuring the stability of topics and topic models, as well as categorizing approaches to improve topic modeling techniques regarding stability issue. Doing so can support researchers who use topic modeling in general, by helping them choose appropriate stability measures, as well as those who focus specifically on stability issues in topic modeling techniques, by organizing existing literature and suggesting future directions.

To accomplish this goal, we conducted a *systematic review* [40]. According to Grant and Booth [40], a systematic review provides a comprehensive fusion of evidence from prior research. Such a review surveys what is known for a given topic and provides recommendation for practice (i.e., Section 5 and 6). It also touches on any uncertainty around prior findings and suggests future research directions for investigation (i.e., Section 7).

To select prior papers for this review we adopt the PRISMA guidelines [76, 87]. This section describes the methodology of the review process, including the research questions the review sought to address, how we

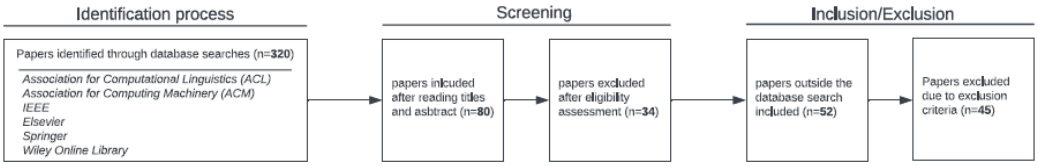

Fig. 4. Overview of this paper's literature review methodology that is adopted from PRISMA guidelines [76, 87]. The process includes three main steps: identification (Section 4.2), screening (Section 4.3), and inclusion based on citations (Section 4.4). The parentheses in each step show the number of total papers obtained or remaining at the end of that step.

identified potentially relevant papers, the process and criteria for screening those potentially relevant papers for inclusion, and the use of citations to identify additional papers. An overview of the process is shown in Figure 4.

4.1 Research Questions

As discussed in Section 2, probabilistic topic modeling techniques generate outputs that vary slightly due to different reasons, e.g., random initialization. This issue of stability has been mentioned in prior work [11, 28, 34, 53, 69, 97], and many different approaches were introduced to measure stability [cf. 51, 72, 81], each with slightly different interpretations. However, these approaches to measuring stability have neither been studied comprehensively nor been organized into different categories (as pointed out in Section 3). This point leads to the first research question of this study. Section 5.1 investigates this RQ.

RQ1*: What are the main strategies to compute stability regarding topic similarity?*

Furthermore, these measures differ not only in how they are computed but also in what information they need to measure similarity in capturing stability. For instance, some focus on word probabilities within each topic, while others focus on topic proportions within each topic.

RQ2*: What information does each technique need to compute the similarity of topics?*

Section 5.2 and 5.3 review different techniques and what they need to measure stability. This review also examines the approaches that are introduced to improve the stability of topic modeling. Generally, each technique focuses on one aspect that causes instability, which leads to RQ3.

RQ3*: What are the categories to divide the contribution of topic modeling techniques that are introduced to improve stability?*

Section 6 addresses RQ3 and reviews different types of approaches developed to improve the stability of topic modeling techniques. In doing so, it provides a conceptual organization that enables meaningful comparisons and contrasts across these different techniques.

4.2 Identification Process

Although topic modeling techniques and applications have been published in a wide variety of journals and conferences, searching for papers in every possible online database would be highly impractical. Specifically, such an approach would likely generate numerous irrelevant results, and would be less suitable given the technical focus of this review. Thus, the process of identifying potential papers for this review started with searching all relevant papers in the following online databases known to the authors for publishing technical topic modeling work: *Association for Computational Linguistics (ACL), Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE), Elsevier, Springer,* and *Wiley Online Library*. To ensure that the list of papers reviewed was not unnecessarily limited by these databases, another round of inclusion was done, as

described later in Section 4.4. Additionally, we did not limit the search process to any specific years to ensure a comprehensive exploration of topic modeling stability in these databases.

In order to search through the aforementioned databases , we used a group of keywords; *stability*, *topic similarity*, and *topic modeling*. Although *topic similarity* may yield different meanings and applications, since similarity of topics across multiple runs of the same model is the crux of stability [cf. 11, 41, 61, 71, 72], including this term when searching in the context of topic modeling or stability increases the chances of finding papers related to stability issue and measures. Next, in the resulting papers, we looked at the abstract or title of the papers, and selected papers that included two or three of these keywords. We made this criteria of inclusion since only including one keyword, e.g., *topic modeling*, would yield numerous papers that did not pertain at all to stability in the context of topic modeling. Additionally, through the search process we found that sometimes *variability*, *robustness*, and *generality* were used alongside topic modeling to refer to topic modeling stability. Thus, we added these words to our keywords. This search process yielded an initial set of 320 papers.

4.3 Screening Process and Criteria

We screened the set of papers obtained from the above identification process by reading the abstract and title for each of those papers. A paper was excluded during the screening process if it was not related to topic similarity or topic stability. For example, a paper that investigated public agenda stability using topic modeling [54] was not relevant to the focus of this work, even though it includes both *topic* and *stability* in its abstract. After this step, 80 papers remained.

Next, an eligibility assessment of papers was done. Specifically, this review was intended only to include prior work that introduced a new method to improve stability of current topic modeling techniques, or work introducing new measure(s) to capture stability. Thus, during this step, papers were excluded if they only employed prior techniques or measures to compare or capture stability [e.g., 2]. Doing so resulted in 34 papers.

4.4 Inclusion of Additional Papers Based on Citations

As discussed above in Section 4.2, including all the publication venues with one or more topic modeling papers regarding stability was impractical due to the high variety of applications. To fill this gap, a round of inclusion was done to find relevant papers that were cited in the 34 papers identified above. In this process, body text of these papers was investigated for other similar works focusing on stability. Similarly, we searched for other papers that cited the 34 papers identified above.

This process identified 18 more potentially relevant papers. However, 7 of those papers were removed, because the language of the paper was not English. At the end of this process, 45 papers were included in the final corpus [1, 5, 6, 11, 20, 21, 24–26, 28, 31, 32, 34, 38, 41, 50, 51, 53, 55, 60, 62, 64, 68, 72, 74, 81, 85, 93, 97, 99, 100, 102, 103, 106, 113, 119, 122, 125–131].

5 COMPUTING STABILITY

As discussed in previous sections, running a topic model multiple times results in different solutions. A highly stable topic model is able to reproduce similar topics across multiple runs with different random initializations and with the same number of topics. This general formulation allows for a variety of approaches to computing stability. These approaches vary in two primary ways. First, which topics across multiple runs are matched for comparison with one another? Broadly speaking, topics can be matched in what this paper refers to as either a pairwise or a recurrent fashion. Briefly, *pairwise* methods make one-to-one unique matches of topics in one run to topics in another run. In contrast, *recurrent* methods allow for one-to-many topic matches by comparing one topic of one run with all topics of another run. Advantages and disadvantages of both methods are discussed in detail in the subsection below.

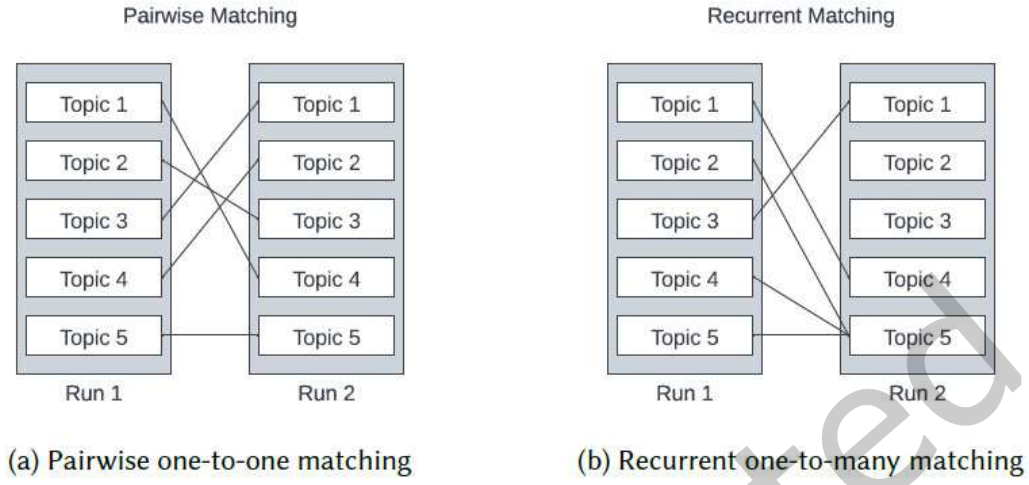

Fig. 5. Different topic matching approaches in computing topic stability. (a) Pairwise techniques match topics of two runs to maximize the average similarity across all one-to-one matches. Each topic in the source run matches with a unique topic in the target run. These approaches allow for locally sub-optimal matches for a single topic in order to maximize the global average similarity across all pairs of topics. (b) Recurrent approaches match topics of a source run to topics of a target run to provide maximum similarity. Multiple topics from a source run are allowed to match to a single topic of a target run in order to maximize the similarity of each pair.

Second, given one topic, how is the topic's similarity to other topics calculated in order to compute the stability of that topic? Similarity can be measured either at the document level or at the term level. At the *term level*, the terms that are assigned to two topics are compared. On the other hand, at the *document level*, document-topic distribution is used to compute topic similarity. Within term level and document level approaches, one can further divide similarity measures into distribution-based or appearance-based metrics. Distribution-based metrics use topic-term probability or document-topic distributions for calculating similarity. On the other hand, appearance-based metrics usually consider a limited number of the top-N terms (or documents) to compute similarity. The rest of this section reviews common stability measures organized by these categories.

5.1 Pairwise vs. Recurrent Approaches

Pairwise approaches make one-to-one unique matches of topics from two different runs for the same model (Figure 5a). Topics are paired using a similarity metric (e.g., matching topics with most similar top terms). Then, average similarity across all pairs is computed to calculate the stability of a model [11, 72]. Usually in pairwise methods, one set of topics (i.e., one run) is considered as a base run, and topics from other runs are matched with topics from the base run [e.g. 72]. Alternatively, pairwise stability can be done by computing similarity of all pairs of runs [e.g. 11]. Computing similarities of all pairs takes K^2 comparisons, while comparing a base run to other runs needs K comparisons.

Recurrent approaches, on the other hand, calculate stability by measuring how often each topic recurs across multiple runs (Figure 5b). Topics of a base run are compared to topics of other runs to see if they match w.r.t. some criteria. The criteria for determining when a given topic has recurred depends on the means used to measure stability, e.g., meeting a set number of repeated terms or documents within a defined threshold [1, 11], or using the similarity of topic-term or document-topic matrices.

Applying either a pairwise or a recurrent approach each has its own advantages and disadvantages. Pairwise methods find the best topic pairs (highest average of topic matches) once, while recurrent methods may count overlapped topics multiple times regardless of any defined threshold. For example, if one topic modeling run has a mixed topic† (topic 5 in run 2 of Figure 5b), and another run has three separated but overlapping topics with that one (topics 2,4, and 5 in run 1 of Figure 5b), pairwise methods only find one highly similar pair. Then, a pairwise method matches other overlapping topics of the second run with a topic of first run that is less similar or even not relevant. Therefore, pairwise methods do not multiple matches and enforce a match in every topic. On the other hand, recurrent methods allow topics to remain unpaired (e.g., Topic 2 and 3 in Figure 5b).

One key factor in computing stability is to avoid conflating models that generate highly stable topics across multiple runs with models that generate highly general topics. Such high generality [55, 85] happens if one or more topics within a single run share similar top terms with different orders and probabilities. Put differently, generality occurs when multiple topics within the same run are highly similar to one another. Recurrent methods for computing stability are sensitive to high generality. That is, recurrent methods may indicate that a model has high stability, but the topics the model generates may not be unique from one another. Pairwise methods perform better in handling high generality. Since pairwise methods make unique matches, they prevent a single, highly general topic in one run from matching with multiple topics from another run. This section discusses different similarity approaches of both types.

5.2 Measuring Stability using Topic-Term Representations

A topic can be represented with term probabilities. Generally speaking, this representation can take one of two forms. First, selecting the top n terms with the highest probabilities for each topic is a common way to display topics [61, 71]. These top n terms are usually listed as a (potentially ordered) set, often without the exact probability value. Second, rather than using only the top n terms, a topic can instead be represented as a probability distribution over the entire vocabulary for a corpus. Either representation can be used to compute a similarity metric to measure the similarity of two topics, though different metrics are used for each of these two different types of representations. The following subsections consider similarity metrics than rely only on a fixed number of top terms, followed by metrics that rely on the full probability distribution over an entire vocabulary.

5.2.1 Top Terms Metrics.

Generally, representing topics as a list of top terms is more readable to human users rather than showing the whole vocabulary with probabilities [41, 60, 85]. Therefore, computing stability using these top terms may be more interpretable to human users, and the results may be easier to understand (e.g., fraction of shared top terms). On the other hand, changing cardinality, i.e., the number of top terms included, can influence human perceptions of topics [59, 109] and perhaps the measured stability.

One common top terms similarity approach is Jaccard similarity. Here, the top terms from two topics are treated as sets, and the similarity is the size of the intersection divided by the size of the union for these two sets (Equation 1) [11, 41, 51, 69, 127]. For a more interpretable comparison, Dice's coefficient can be used. It is calculated using twice the size of the intersection divided by the sum of the two sets' sizes (Equation 2) [81]. Since A and B here are top terms for two topics, the denominator ends up being twice the cardinality. And because this formula divides twice the intersection by the sum of cardinality of two runs (while Jaccard uses union length), the value of Dice's coefficient can be interpreted as the exact proportion of the intersection. For example, a Dice's coefficient of 0.90 means 90% of all terms are similar in both topics, while there is no comparable direct interpretation for a Jaccard's similarity of 0.9.

† A mixed topic includes top terms from two or more topics that are represented either in the same or in different runs or models [18]

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$Dice = \frac{2 * |A \cap B|}{|A| + |B|} \tag{2}$$

Alternatively, to increase sensitivity of positions of top terms, rank correlation between top terms can be used to compute similarity of topics. Similar top terms with similar positions in two sets show a stronger similarity. A rank correlation similarity approach penalizes similar terms with different positions and increases the penalty as position difference increases. One such metric is Spearman rank correlation that computes similarity using multiple rank variables [130]. One variable captures the order of the terms from 1 to n. Another variable is ranked to capture whether a term is in both topics or not (e.g, 1 for being in both and 2 as a lower rank for being exclusively in one topic). Mantyla et al. [69] used Spearman rank correlation to compute stability of multiple runs. However, the authors did not state how they exactly ranked and penalized exclusive terms.

Although, Spearman rank correlation accounts for term orders in computing similarity, this approach is not able to distinguish between similar topics with different term orders and topics with no shared terms. Spearman correlation of two topics with same set of words but opposite order is -1, even though one may argue that these topics are not completely opposite. On the other hand, Spearman correlation of two topics with no intersection (i.e., no top terms in common) is -0.86. Spearman correlation rank is a suitable metric for monotonic comparison (if one vector is a monotone function of the other), which is different than computing term similarity of shared terms. In contrast, as an advantage, considering term ranks prevents the stability computation from matching topics with shared terms but different orders.

Any of these three metrics introduced above(Jaccard similarity, Dice's coefficient, and Spearman correlation) can be used to compute stability via both pairwise [41] and recurrent approaches [5, 69, 128].

Generally, stability is computed by averaging across matched topic pairs. Alternatively, Descriptor Set Difference (DSD) approach computes the number of different elements in two sets [11]. The DSD approach, Equation 3, computes the number of different top terms in all topics of two runs ($T_1 \Delta T_2$) and divides it by number of topics times top-N terms. DSD is sensitive to the number of top terms and outputs higher values (i.e., lower similarity) for lower numbers of top terms.

$$DSD = \frac{|T_1 \Delta T_2|}{t * K} \tag{3}$$

In order to remove sensitivity to the number of top terms, Agrawal et al. [1] introduced a new metric, Median of term overlaps, to capture shared terms of two runs across multiple number of top terms. First, each solution (e.g., a run or a model) is represented as a bag of words of top terms of the generated topics. Then, similarity of two runs is computed as the number of shared terms between those two runs. To compute stability, the average of similarity across multiple runs is computed; however, treating all topics of one run as a set of terms, instead of set of topics, does not necessarily imply similar set of topics.

Most of the discussed stability approaches do not consider word order (except Spearman rank correlation). Mantyla et al. [69] used an order-based similarity approach, Rank Based Overlap (RBO), introduced by Webber et al. [123]. Put concisely, RBO computes the similarity of two ranked lists (i.e., top terms of two topics) with a weighting schemes that controls the importance of term ordering and length of the lists (number of top terms).

As mentioned earlier, topics are usually represented as top terms to human users. That does not necessarily mean similarity metrics based on top terms provide more accurate similarity and stability comparisons. For that reason, this section described common top terms similarity and stability methods and discussed advantages and disadvantages each metric offers. All top term metrics are sensitive to the number of top terms. It is not clear which number of top terms can provide higher accuracy in stability comparisons. Besides, most of these

methods (except RBO) ignores the term order, while different orders may represent different topics. Therefore, in choosing one metric over another, selecting number of top terms and considering the order of those terms are two important factors.

5.2.2 Term Probability-Based Metrics.

Using topic-term probabilities is an alternative to computing stability using top terms. Choosing N to compare top-N is a crucial point that affect the stability computation. Besides, metrics based on top terms are essentially ignoring information about the topics. Looking only at which terms are included in the top-N, and possibly term order of the top-N terms, ignores the detailed information of topic-term probability. On the other hand, using term probabilities, a similarity metric can capture differences between two topics with shared top terms but different probability magnitudes.

Cosine similarity computes the angle between two vectors with same size. Equation 4 shows how cosine similarity is calculated for two vectors A and B in which "·" indicates dot product and "||A||" stands for magnitude of vector A. Cosine similarity bounds between -1 (completely opposite vectors) and +1 (thoroughly similar) [72, 81]. One disadvantage of cosine similarity is that this metric is more sensitive to pairs' product with higher values. It produces very small values (close to zero) for most of topic-term probabilities. Thus, top terms with higher probabilities affect the final similarity results more than other terms.

$$Cosine(A, B) = \frac{A.B}{||A|| * ||B||} \tag{4}$$

Another common stability measure, Kullback-Leibler divergence, computes distance of two partitions [53]. Equation 5 shows KL divergence calculation for two topics P and Q across two runs with T words (T=Set of all words in the vocabulary). Unlike cosine similarity, KL distance considers ratio of differences using $log(P/Q)$. In order to use KL distance for similarity, it should be normalized to [0,1] and subtracted from 1 [52].

$$KL(P||Q) = \sum_{x \in T} P(x) * log(\frac{P(x)}{Q(x)}) \tag{5}$$

KL divergence is not a symmetric measure, i.e., $KL(P||Q) \neq KL(Q||P)$. Instead, Jensen-Shannon divergence (JSD) similarity metric computes KL divergence for both P and Q w.r.t. to a defined middle point called M. JSD is a symmetric measure (Equation 6) [51, 72, 81]. JSD metric values are normalized between [0,1], but still needed to be subtracted from 1. Although a symmetric KL divergence measure is also defined and is averaged across both $KL(P||Q)$ and $KL(Q||P)$, in topic similarity applications, asymmetric KL divergence and JSD is more common.

$$\begin{aligned} JSD(P||Q) &= \frac{1}{2} * KL(P||M) + \frac{1}{2} * KL(Q||M) \\ M &= \frac{1}{2} * (P + Q) \end{aligned} \tag{6}$$

These three metrics described above capture similarity of topics across multiple runs and can be used to measure stability in both pairwise and recurrent approaches. For example, Miller and McCoy [72] introduced topic proportion alignment to compute pairwise stability using any of these three metrics. This approach calculates the number of aligned topics (w.r.t. a defined threshold for any similarity measure) over average number of topics for two runs (can extend to more than two runs/models). They also introduced weighted similarity which is an extension of cosine similarity that adds higher weights for topics that are repeated more across multiple runs.

Sections 5.2.1 and 5.2.2 discussed stability computation using term similarity. Another category of topic similarity is discussed in the next section within document point of view.

5.3 Measuring Stability with Document Similarity

As an alternative to computing stability using topic-term assignment, one can instead find similar topics based on document-topic assignment. As reviewed in section 5.2, in measuring stability with term similarity, a highly stable topic model is able to generate similar topics with similar top terms or topic-term probabilities across multiple runs with different initialization. Similarly, running a highly stable model multiple times should also return similar document distributions with small variations.

One advantage of using document similarity measures is that the relationship between documents is more reliable than the word relationship; that is the semantically similar words may be assigned different probability values but does not affect the relationship between representative documents of a topic widely [28]. Additionally, a group of similar documents are more likely to form a similar topic with slight differences (e.g., different orders or distribution values), while similar set of words with different order can be perceived differently (e.g., by human assessors) for a topic [45]. Any vector similarity metrics (e.g., cosine, KL, JSD) can be also used to compute similarity using document-topic assignment. The rest of this section reviews metrics that were initially or specifically used to measure stability within document similarity.

De Waal and Barnard [28] introduced a metric to compute stability using document similarity. As an example, suppose there is topic A from run r_1 of a given model and topic B from run r_2 of that same model. De Waal and Barnard designed a weighting scheme to compute the product of document distributions of the two topics as a graph's edge weights (Equation 7).

$$Sim(A, B) = P(T_A|D) * P(T_B|D) \tag{7}$$

After computing these similarities for topic A from r_1 and topic B from r_2, it finds the optimal matches using Hungarian method [38] and the obtained graph's edges. Thus, each topic in run r_1 aligns with a topic of run r_2 of the same model with similar settings [28]. Similar to the cosine similarity metric, small distribution values produce very small outputs and make the metric sensitive to document-topic distribution with higher values.

Finding similar topics can be done by looking at the top documents of each topic across multiple runs. Belford et al. [11] applied Normalized Mutual Information (NMI) to topic-document assignment of two runs to compute topic similarity [114]. In this approach, a document is assigned to one dominant topic using the document-topic distribution matrix. Overall level of agreement between all topics then is computed and averaged across multiple runs (called Pairwise Normalized Mutual Information) as the stability value. Considering only the dominant topic of a document makes this metric more sensitive to different document orders and initialization. On the other hand, high values of PNMI's stability across multiple runs for one model shows the model has high reproducibility and stability from document point of view[†].

5.4 Summary

Table 1 summarizes the stability metrics reviewed in this section. For each technique, this table summarizes whether the technique can be used for different means of selecting topics to compare (pairwise or recurrent), the level at which the technique can be applied (term or document), and the type of data that each technique uses for computing similarity (appearance of top terms or documents, probabilities of individual terms or documents, or probability distributions over multiple terms or documents). Thus, Table 1 addresses both RQ1 (i.e., what the

[†]Stability and reproducibility are not necessarily the same. Some of stability approaches (especially recurrent approaches) does not compare pairs of topics. These methods compute similarity of top documents or terms of different runs in overall. Thus, high generality could be the cause of high stability, but it is possible to have different topics and low reproducibility at the same time. High reproducibility and stability at the same time leads a model to generate similar topics at each run. In contrast, high stability and generality may make a model to generate similar shared documents or terms but with different topic solutions.

Method	Comparison		Level		Similarity	
	Pairwise	Recurrent	Term	Document	Appearance	Probabilities
Cosine [72, 81]	✓	✓	✓	✓	✗	✓
KL [53]	✓	✓	✓	✓	✗	✓
JSD [51, 72, 81]	✓	✓	✓	✓	✗	✓
Proportion alignment [72]	✓	✗	✓	✓	✗	✓
Jaccard [11, 41, 51, 69, 127]	✓	✓	✓	✗	✓	✗
Dice [81]	✓	✓	✓	✗	✓	✗
Spearman [69]	✓	✓	✓	✗	✓	✗
DSD [11]	✗	✓	✓	✗	✓	✗
Term overlap median [1]	✗	✓	✓	✗	✓	✗
RBO [69]	✓	✓	✓	✗	✓	✗
Doc weighting scheme [28]	✓	✓	✓	✓	✗	✓
NMI [28]	✓	✓	✓	✓	✗	✓

Table 1. Similarity and stability techniques comparisons. Techniques and their references are listed under the Method column. The Comparison columns indicates how topics of different runs are compared. The Level columns indicates whether each metric can be used on terms, on documents, or on both levels. The Similarity columns show what data are used to compute similarity: top-terms (appearance), topic-term probability, or document-topic distribution.

main strategies are to compute topic stability) and RQ2 (i.e., what information is necessary to compute different measures of topic stability).

Additionally, Figure 6 summarizes Table 2 further and shows the distribution of the similarity and stability techniques for each category (i.e., Comparison, Level, and Similarity) divided by their associated aspects. This helps to understand what proportion of the stability techniques were focused on each aspect of a category. For example, the last row of this figure, Similarity, indicates that half of the techniques measure stability based on appearance (i.e., term appearance) and the rest half use term probability or document distribution to measure stability of a topic model.

Selecting one similarity metric and stability approach over another seems difficult with each method having different merits and demerits. Selection of a stability metric should be guided, first, by the nature of the data being analyzed and, second, by the ways that a stability results will be used (e.g., assessing model quality vs. visualization). For example, if the dataset that is being used includes less diverse latent themes (and if the researchers are aware of such prior knowledge) and topics shared high number of top terms, pairwise methods should be selected due to their robustness to high generality. On the other hand, cardinality is an important factor in choosing one method over another. Low cardinality of top terms causes lower generality of topics (since top terms are more unique and less repetitive); therefore, recurrent approaches are able to capture similarities and differences more accurately, while pairwise methods may miss mixed topics. In addition, top terms stability metrics are more appropriate for visualization purposes compared to probability based metrics. This is because, in visualizing topics, stability of top terms is also important for human users. A researcher may also want to select a metric at term or document level. As a comparison example, a term stability metric is a proper choice for

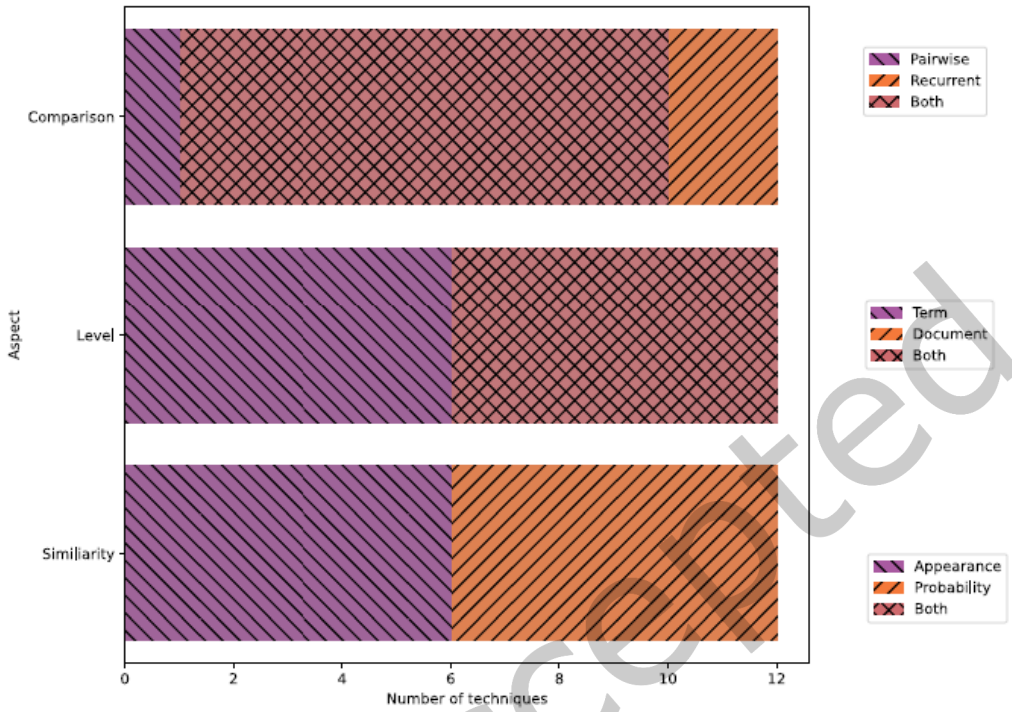

Fig. 6. Distribution of similarity and stability techniques regarding different categories; i.e., Comparison, Level, or Similarity. Each row shows number of techniques, sums up to 12, that can be used to measure one or more aspect of a category (e.g., Pairwise, Recurrent, or both aspects for Comparison category). We see that the majority of techniques can be used for both pairwise and recurrent comparisons; that all techniques can be used at the term level, while there are no techniques specific to the document level; and that similarity measures either use appearance-based metrics or probability-based metrics, but never both.

labeling methods, while document level measures are more suited for summarizing topics within documents. Needless to say, a more comprehensive comparison can be done by using multiple metrics.

6 TOPIC MODELING AND STABILITY

Just as different metrics have been devised for assessing stability, different techniques have also been developed to improve stability in topic modeling. Some techniques work by modifying different parts of existing topic models, while others work by introducing entirely new models.

De Waal and Barnard [28] were among the first researchers who suggested using stability to evaluate a topic model. They introduced a document weighting scheme to compute document-topic correlation in order to measure topic stability (Table 1). While their work did not introduce a new topic modeling to improve stability, it was an early work that suggest stability may vary among different topic modeling techniques.

This section reviews different approaches to improve stability of topic modeling, including: different topic models, alternate generative process (i.e., sampling and/or inference algorithms), combinations of existing models, use of multiple random initialization, and (hyper)parameters adjustments. Although improvements in topic

quality are not the main focus of this paper, approaches with quality enhancement that also improved topic stability are discussed to demonstrate how topic stability can relate with modifications intended to improve quality.

6.1 Novel Models

This subsection compares and contrasts novel topic models (different probabilistic approaches, matrix factorization methods, etc.) against previous common models discussed in Section 2. Delving into all alternatives models would exceed the scope of this paper. Instead, this section focuses only on models that were specifically intended to improve stability.

An early example comes from De Waal and Barnard [28], who computed stability for LDA and for a matrix factorization method. They used Gamma-Poisson (GaP) as the matrix factorization method, which chooses topic proportion of a document with Gamma distribution. GaP updates topic-term matrix for each word of the current document using Poisson distribution to reduce reproduced documents' terms error [20]. De Waal and Barnard [28] showed that the LDA model is able to produce more stable results across two runs, while a single run of GaP generates topics that are less similar to other topics within that run (i.e., the topics are more diverse). While this work does not introduce a novel method to improve stability (GaP factorization had already been introduced for topic modeling [20]), it compares topic modeling techniques from a novel standpoint that may function differently than other topic metrics. The relationships between stability and other metrics are discussed in more details in Section 7.

Beside Probabilistic topic models (e.g., pLSA, GaP, and LDA) and matrix factorization techniques (e.g., NMF, and SVD) that attempt to reduce document-term generation error using an optimization process, there is another type of topic models that cluster terms or documents as topics. These models use a similarity-based metric to cluster topics and represent each topic's descriptors as their highly important terms (for example high ranked TF-IDF terms of each cluster). The rest of this section is organized to describe the most recent similarity-based methods and discuss these models' stability performance.

El-Assady et al. [31] introduced a new hierarchical document/topic clustering method called Incremental Hierarchical Topic Modeling (IHTM). This method generates topics by clustering documents using document similarity. Document similarity is computed within terms of the documents (e.g., TF-IDF or other term-frequency features). Then two or more similar documents form a topic. Document's keywords (e.g., top-N words with the highest TF-IDF) are merged to generate term representation of a topic. Originally, IHTM was designed to improve visual representation of hierarchical relationships among topics and to provide parameters (e.g., threshold for merging topics) for user interactions.

However, it has also been shown that IHTM also improves stability. El-Assady et al. [32] used a newly defined metric called Document-Matched-Frequency (DMF) to compare LDA, IHTM, and NMF. DMF captures the similarity of a topic's top documents with a set of ground truth documents, which differs from computing stability by comparing results across multiple runs.†

IHTM resolves the issue of document order wherein providing the same input documents in a different order can yield varied topic solutions (discussed in Sections 1 and 5). However, IHTM is only able to resolve this issue for document-topic assignment. In topic-term assignment, IHTM is highly sensitive to document order. This occurs because a topic's descriptors (i.e., assignment of topic top terms) are formed based on incoming documents, and latter documents have less influence on descriptors compared with early ones. Moreover, this model is not able to provide document-topic distribution or topic-term probability. This makes IHTM more suitable for certain

†Since DMF requires ground truth documents, a highly unusual requirement not found in any other stability metric, we omit DMF from the survey of stability metrics in Section 5.

visualizations, but not suitable for all purpose as the comprehensive probability distributions generated by other topic models (e.g., LDA).

6.2 Improving Topic Generation and Inference

The inference phase of topic generation is an unsupervised process and one source of instability in topic modeling. For example in LDA, the optimization procedure depends on drawing latent parameters (e.g., topic-term β) before the generative process starts. These latent parameters determine document-topic proportion (θ) and topic assignment of words (z_n), which are also not observed. Altering the procedures by which these latent parameters are inferred may be able to improve stability.

Roberts et al. [98] introduced a new approach to improve LDA optimization by making an initial guess for θ and z_n and then optimizing β with Variational Expectation Maximization (VEM). This approach is called Structural Topic Modeling (STM). To make more accurate guesses, Roberts et al. [97] added two covariates, topical prevalence and topical content. The prevalence covariate is based on an assumption of a linear relation between a document discussing a topic and metadata variables (e.g., for political blog posts, whether the author is liberal and conservative). The content covariate takes documents' different sources (e.g., blogs vs. news media, or different news media venues) into account in the generative process. In the inference phase, θ is drawn from a logistic-normal distribution of topical prevalence (X) and topic covariance (Σ).

STM is built to resolve LDA's optimization issue in the presence of latent parameters. LDA optimization depends on starting points in finding local minima [98]. STM resolves this dependency on the starting point by adding covariates, topic covariance, VEM, and an initial guess (discussed in more details in Section 6.4) to the optimization procedure.

Although Roberts et al. [98] did not compare the stability of LDA and STM across multiple runs, they compared the stability of the relationship between the covariates and estimation of topics proportion (θ_d). They observed that STM relationships of generated topics to covariates are more stable than LDA generated topics. To estimate topic proportion (θ_d), MAP[†] was used which adds priors to likelihood computations [98, 100]. The authors used MAP analysis to compare LDA, STM, and the true distribution of a synthesized data set. This comparison showed which model is more stable in capturing covariate-topic relations. Figure 7 shows STM can capture covariate's effect that LDA is not able to integrate. Besides, running STM multiple times does not affect covariate-MAP widely from true synthesize data distribution. This shows, besides using VEM algorithm, adding covariates provides stable inference phase and β estimate for STM. However, since LDA does not integrate covariates to initialize and optimize the parameters, this comparison does not necessarily mean one is more stable than the other in generating final topics.

As another attempt to resolve the topic generation instability issue, Koltcov et al. [53] introduced a new sampling approach that they term granulated LDA (gLDA). Generally, topic models use a bag-of-words representation that does not account for the proximity of neighboring terms within a document either in the generative process or, more specifically, in the sampling procedure. The authors introduced a new sampling method called Granulated Gibbs sampling that samples terms randomly similar to Gibbs sampling, but assigns same topic to the neighbor terms around anchor words (top terms derived from previous inference iteration) with a fixed window length. It has been shown that increasing the length of the neighboring window increased stability of the topics generated by gLDA according to Jaccard stability measure. Koltcov et al. [53] ran LDA (with Gibbs sampling and variational Bayes inference), pLSA, and gLDA 3 times on the same corpus. The results showed that gLDA increased stability according both to Jaccard similarity and to symmetric KL. Interestingly, gLDA also increased topic quality according to coherence [75] and tf-idf coherence [82].

[†]Maximum of a Posteriori

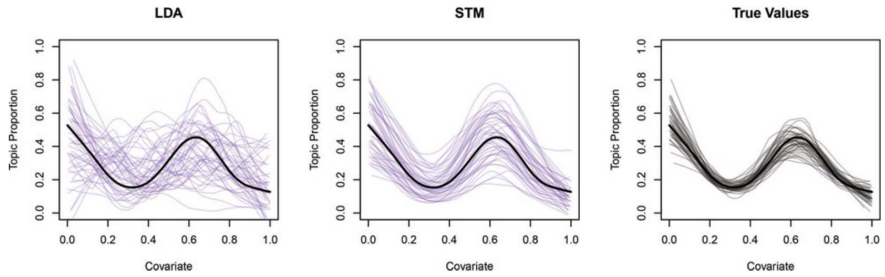

Fig. 7. Covariate-topic relation fitted plot on 50 simulated text datasets [97]. Each line shows the relationship between topic proportion of a topic model and defined covariates. The True Values plot displays the true relationships of covariates and topics in each of 50 synthesized datasets. The LDA plot indicates this model is not able to capture true relationship and perform with low stability. The STM plot is closer to the true distribution of the data with little changes across multiple runs.

Koltcov et al. [53] and Roberts et al. [97] used two different inference algorithms, sampling-based (gLDA) and Variational Expectation Maximization algorithm (VEM), and both improved stability. However, Yao et al. [129] showed different inference algorithms affect topic solutions by comparing different Gibbs sampling (which gLDA originally derived from), VEM (which STM used for optimization), Max-Entropy, and Naive Bayes. Since the true distribution of topics is not observable, the authors used a version of Gibbs sampling that jointly resamples all topics (computationally expensive to be used regularly) to generate document samples. This posterior of topic distribution is assumed to be closer to the true posterior since it jointly resamples all topics. This comparison emphasizes that none of the topic modeling algorithms can produce results that are similar to what authors referred as a closer true distribution and compared other models against w.r.t. F1, cosine similarity, and KL-distance metrics. Although Yao et al. did not compare the stability of different models across multiple runs, comparing similarity to a base model with the ability to update all parameters at once reveals that altering inference algorithms affects topic models' ability to generate solutions/topics. That is, using different sampling and optimization techniques may improve stability, but the resulting final solutions may also be less similar to a synthesized "true" distribution.

6.3 Combining Multiple Runs' Solutions

In this and previous sections, models that altered the topic modeling algorithm and/or inference phase were reviewed. Besides changing algorithm of a model, combining multiple runs is another way of improving topics stability. Next section discusses advantages and disadvantages of such combination methods. Rather than attempt to alter the underlying model or inference algorithm, an alternative approach to achieve higher stability in topic modeling is to combine topics of multiple runs. With each subsequent run, each topic is either a new topic or repeated topic from a previous run, as determined by similarity measures (introduced in Section 5). Merging repeated topics and adding new, distinct, non-repeated topics is expected to produce a stable solution. Combining multiple runs may end with different number of emerged topics than the initial number of topics (e.g., caused by different settings, different number of initial K, or changing similarity threshold for merging). Different approaches, discussed further below, handle this issue differently.

Belford et al. [11] introduced a new topic modeling ensemble that combines topics of multiple runs. The method proposed by Belford et al. [11] consists of two steps: generation and integration. In the generation step, multiple number, termed r, of NMF models are executed, each generating K topics. These $r * K$ topics are stacked vertically to form a topic-term matrix M. Then, in the integration step, an NMF decomposition algorithm is applied to

matrix M to generate topic-term matrix H with k components (topics). Topic-term matrix H is multiplied by the original document-term matrix and produces document-topic matrix D. The authors used Non-Negative Double Singular Value Decomposition (NNDSVD) algorithm (an NMF technique) in the integration step that differs from least square NMF used in the generation step. Boutsidis and Gallopoulos [17] found that initializing factors of an NMF technique (e.g., SVD, least square) with an SVD has a huge impact on final solutions and reduces random initialization sensitivity. This method is called Non-Negative Double Singular Value Decomposition (NNDSVD).

Authors compared stability of LDA and NMF topic models using DSD and average Jaccard similarity (introduced in Section 5). Despite the higher stability of the ensemble model, the document-topic matrix is generated after the model finishes matrix decomposition (not within either of the two steps). This may cause different document-topic solutions each time the ensemble model is executed, since the r topic models (and thus M matrix) vary at each run. Belford et al. [11] did not provide any stability comparison at document level to investigate possible effects of the ensemble strategy on the generated document-topic matrix.

In short, ensemble of NMF topic models can achieve higher stability than a single run of LDA or NMF model. However, running a matrix factorization method on the results of previously obtained factors makes the interpretation difficult. Alternatively, clustering to combine topics of multiple runs generates more interpretable results.

Mantyla et al. [69] introduced a new method to cluster generated topics of N runs of LDA models. This method employed K-medoids clustering, which starts with a selection of cluster centers (either randomly or with an approach to minimize an objective function) and assigns data points to each center. K-medoids iterativly searches for the best replacements of cluster centers. Replacements are selected from each cluster's data points to reduce an intra-cluster distance measure. At each stage centers and assigned data points are updated until no more changes occur [89, 105]. After obtaining t=N*K topics (N runs, each with K topics) with topic-term matrix of size t*w, authors employed the GloVE word embedding technique [91] to project this matrix to a lower dimension matrix of size t*v ($v = [200, 400]$, $v < w$).

Running K-medoids with K (i.e., the same as number of topics) clusters results in generating K topics. The proposed method sums up the term probabilities of topics of each cluster and selects top_{10} terms as the descriptors of a cluster. Because of this combination of topics, a topic-term probability matrix is not provided. Authors did not compare LDA clustered stability with LDA or any other topic models, but used clustering metrics (e.g., Silhouette) and stability metrics (e.g., Jaccard) to show the best, the median, and the worst topic of 20 runs of this model according to these metrics.

Mantyla et al. [69] uses word embedding in clustering topics, thus topic stability is subject to change with different word embedding training parameters, trained corpora, initialization, or document order [5]. Moreover, no topic quality assessment and comparison is provided for this method, and it is hard to analyze how this approach to stability affects usability of the results or their perceived quality.

Clustering topics may combine mixed topics[†] with non-mixed ones. Whether this occurs depends massively on the pre-defined number of clusters and distance measure. A hierarchical clustering stops merging topics that are partially similar at lower levels and combines them as it goes to higher level clusters. Miller and McCoy [72] proposed a new hierarchical clustering model to improve stability of topic modeling and hierarchical summarization. First, the model executes topic modeling under the same conditions (e.g., same number of topics, documents, and pre-processing) multiple times. Second, it finds pairwise matched topics of different models using cosine similarity. Third, it clusters topics by applying group agglomerative clustering [33] using computed pairs' similarity. Fourth, for each cluster, it (I) assigns cluster's members, (II) updates centroids using hierarchical topic models (ignore topics with less than a threshold similarity to centroids), and (III) computes clusters' stability of updated centorids. This clustering methods repeats these four steps until no changes happen.

[†]See footnote †.

In this approach, Miller and McCoy [72] used Hierarchical Dirichlet process (HDP) with Gibbs sampler [118] as flat topic model (first stage), and nested DHP [88] for hierarchical clustering (stage 4.II). Three stability metrics were used to capture similarity of topics across multiple runs: Cosine similarity, ratio of aligned topics, and JSD. This hierarchical topic modeling was not compared to any other model, but it has been shown with deeper hierarchical alignment of topic models, stability decreases Miller and McCoy [72]. Although this model tends to improve stability, the results are subject to change with different ordering, especially since centroids are shaped based on topics' order of entry into the clusters.

These last two methods [69, 72] used top terms' similarity to merge topics and ignored document similarity. El-Assady et al. [32] suggested using both terms and documents similarity to match and merge topics based on an observational study of manual matching. Authors defined a strong or complete match if a match maximizes both term and document similarity between topic pairs of two different runs. LTMA (Layered Topic Matching Algorithm) model was introduced to do so. It first calculates topic matching and generates matching candidates. Then, it adds candidates if they are complete-matches as higher ranks, or document-only and term-only similarity matches as lower ranks. They used DMF as described in Section 6.1 to capture similarity of topics at the document level. A new weighted term similarity measure called Ranked and Weighted Penalty Distance (rwpd) is introduced. The rwpd similarity measure puts more weight on similar higher ranked top terms and penalizes missing top terms with higher ranks as well. A comparison between LTMA and LDA showed LTMA is more stable within DMF metric. However, authors did not compare stability of previously introduced IHTM (a hierarchical matching model) [31] with LTMA that both were showed to be more stable than LDA.

Considering both term and document similarity is a plus in an ensemble model. In contrast, LTMA (similar to IHTM) is not able to produce topic-term and document-topic matrices. Besides, neither of these two models, IHTM and LTMA, can be used as an ensemble base model. Moreover, for quality assessment, authors compared LDA, IHTM, and LTMA according to the amount of correct document-theme assignment based on a defined ground truth. LTMA performed better in assigning similar documents of a theme to a topic. This shows LTMA was able to improve stability and quality of topics. However, using this measure and comparing document-topic assignment with a pre-defined document-theme might not be the best way to compare stability or quality of topics. Although topic models are supposed to extract latent themes, that does not necessarily mean extracted themes and defined themes will match thoroughly. Besides, generating different topics compared to ground truth themes does not indicate whether the model is or is not stable.

Section 6.1 through 6.3 reviews how changes in a model affects stability if those changes modify the whole models, alter the inference phase, or combine multiple solutions. Since most of topic models are sensitive to initialization and pre-defined parameters to obtain topics, the next two sections analyzes these small changes and their influence on topic stability.

6.4 Initialization

Changing parameters and settings of a topic model varies the generated topics. A topic model initializes some of these parameters randomly (e.g., topic-term matrix). Different random initialization will yield different topic solutions. Prior work has considered using different initialization approaches to increase stability, as described in this subsection.

As discussed in Section 6.2, STM alters LDA by adding two covariates and other hyperparameters to estimate topic-term and document-topic matrices (β and θ). However, STM still ends up with different topics when starting with different random initializations. Roberts et al. [98] showed that final solutions can become more stable with higher quality (measured negative log likelihood of the model) if LDA or spectral algorithm [6] is used as initialization. Unlike LDA, the spectral algorithm extracts topics with provable guarantees to maximize the likelihood of the objective function. Thus, it is deterministic and stable.

To measure the stability, Roberts et al. computed lower bound of the marginal likelihood of multiple runs. The VEM approach with JSD inequality is used to derive a lower bound of marginal likelihood of the unobserved parameters (topics) given the observed data (documents). They used this measure to compare stability and quality of different initialization techniques. In this comparison, higher values of the lower bound of marginal likelihood [116] yields higher quality and smaller changes of values across multiple runs indicates higher stability results.

Roberts et al. [98] compared these two initialization methods – LDA and spectral – and found that using LDA helps STM converge faster. Thus, it was chosen as the default initialization method of STM's R package [99]. They also found that using more iterations of LDA initialization does not improve the stability and quality of the generated topics. While an iteration of the spectral algorithm is slower than an iteration of LDA, the spectral algorithm converges after only one iteration.

Roberts et al. also stated that using either the LDA or the spectral initialization method causes STM to perform with a lower bound of marginal likelihood (higher quality). It has been shown that STM with pre-initialization achieves higher stability and quality using the lower bound of marginal likelihood. However, quality (e.g., perplexity, or NPMI) and stability (e.g., average Jaccard) of the STM generated topics were not compared to topics of the other models. Furthermore, although STM can employ either an LDA or a spectral initialization for a more stable start, it does not necessarily mean the inference process – including VEM optimization, sampling, and generative process – provides more stable (with higher quality) solutions than LDA at the end. Especially because LDA itself uses random initialization, this may influence the stability of STM results as well.

STM is not the only topic model that employed another topic model to initialize parameters. As previously discussed in Section 6.3, Belford et al. [11] showed that a single NMF model with NNDSVD matrix factorization technique is more stable than an ensemble of least square NMF models. Initialization of multiple models reduced the quality (according to NPMI and NMI) and stability (according to average Jaccard and average DSD) of the proposed ensemble model [11]. Therefore, [11] designed a new ensemble model by breaking the dataset into f folds of documents. So, instead of having r NMF models with random initialization (previously discussed in Section 6.3), they executed multiple NNDSVD topic models, termed p, on each fold.

In the next step, topics are generated similarly to the process described in Section 6.3 for the approach [11] introduced. The proposed f-fold ensemble model performs better than LDA, NMF, and ensemble NMF, but mostly similar to NNDSVD w.r.t. stability and quality comparisons introduced above. Interestingly, the models with the highest quality (NPMI and PMI) in experiments were not necessarily the most stable ones. When LDA or a single model NMF showed higher performance, they were among the least two stable ones. This may indicate that the most stable models are not always the models with highest quality and vice versa. Comparing and contrasting stability and quality is discussed more in details in Section 7.

6.5 (Hyper)Parameter Adjustment

Beside random initialization, (hyper)parameters adjustment can also affect solutions of a topic model [25]. Various hyperparameters optimization techniques have been used to achieve higher stability in topic modeling. Chuang et al. [25] optimized α and β hyperparameters using a parameter exploration technique called Grid search. They show that hyperparameters changes affect resolved and repeated topics and alter fused and fused-repeated topics[†] even more. Agrawal et al. [1] analyzed topic modeling papers and stated that, as of 2018, only one third of the topic modeling papers that focused on stability used some level of parameter tuning (i.e., manual adjustments or

[†]Chuang et al. [25] introduced topical alignment between topic solutions of a model and a topic reference (e.g., topic reference can be derived from domain experts, meta-data, and/or topics of a base model). Authors used expert-authored concepts as topic references for the comparisons. They categorized topics into resolved topics if there is a one-to-one correspondence between a topic of a model and a topic in reference, fused topics if a topic is a mixture of two or more referenced topics, repeated if there is a many-to-one matches of the generated topics and a reference topic, and fused-repeated the many-to-one match is between fused topics and a reference topic.

parameter exploration) and less than 10% of those papers mentioned that parameter adjustments had a huge impact on final results.

To explore this gap in the literature, Agrawal et al. [1] introduced a new evolutionary algorithm (EA) technique to maximize topic modeling stability using hyperparameters adjustments. They employed the Differential Evolution (DE) algorithm†, which they chose because it has been shown that DE is competitive to Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). LDADE starts with randomly initializing N (number of population) sets of parameters including number of topics (K) [10,100], α [0,1] and β [0,1]. Then, the algorithm iteratively generates new sets of parameters and evaluates new offsprings by running an LDA on each set of parameters; authors compared both Gibbs-LDA and VEM-LDA on LDADE evaluation and found out Gibbs-LDA provides more stable topics.

Agrawal et al. [1] then computed the number of similar (overlapped) terms in topic pairs for different number of top-N terms. The computed the ttability of a model using median of overlaps, R_n, across multiple runs. They called this approach raw score, then introduced another measure, Delta score, to compute the difference of R_n before and after DE optimization. Delta score measures the stability improvement or changes. To reduce the effects of document order on stability, Agrawal et al. [1] evaluated each model within 10 runs of LDA using shuffled data.

While it has been shown LDADE improved both stability (median of term overlaps) and quality (F1 score for text classification using generated topics) of LDA topic modeling, it takes almost 5 times longer than a single LDA to run. On the other hand, changing parameters of DE algorithm (e.g., number of iterations, cross-over rate, or differential weights) affects stability and the generated topics. Therefore, LDADE itself needs exploration and adjustment, too. Despite classification improvement, Agrawal et al. [1] did not compare LDA and LDADE using other quality measures (e.g., NPMI, or Perplexity). Moreover, choosing the the best model using LDADE is done by assessing classification via F1 score. This binds the resultant model to be more suitable for the assessed tasks, classification on a specific corpus in this case, rather than the the quality of topics themselves including human perception of quality or a more general topic quality assessment (e.g., Perplexity, or NPMI).

6.6 Summary

This section describes a number of different strategies in altering topic modeling – algorithmic changes, combining models, initialization variations, (hyper)parameters optimization – and each one's influence on stability was discussed. The above description of these strategies addresses RQ3 (i.e., the different categories of contributions to improve topic modeling stability). Table 2 summarizes the prior work focused on improving topic stability. All of the model listed in the Table 2 improved one or two topic stability measures, thus each model may only improve stability regarding to the specific measures and not all the measures listed in Table 1.

At the time of writing this paper, little research has compared stability across different topic models. Such comparisons would be useful to help researchers choose a topic model, to select number of topics to achieve higher stability (if it is desired), or to compromise between other measures (e.g., quality) and stability. Choosing a model may cause the final outputs to vary slightly from the original approach or compared approach to completely different results, including different number of topics, top terms, top documents, or interpretation. Selecting a model should be guided not only by stability measures but also by other factors such as metric evaluation or human assessment. This point is considered further below.

†Similar to other evolutionary algorithms, DE starts with a random population which is called agents. In each iteration of DE, for each agent x, three other distinct random agents are selected. Agent x's new positions randomly remain the same or are created based on a combination of three previously selected agents. If The objective score of the new generated agent is improved, it replaces the old x agent. This algorithms stops when it meets the finishing criterion [113] (e.g., maximum number of iterations, or reaching a pre-defined value).

Type	Technique and/or Citation	Short Description
Novel Models	IHTM [31]	Clustering top terms and top documents
Topic Generation & Inference	STM [100]	Added co-variates & changed optimization process
	glda [53]	Used Granulated Gibbss Sampling
Combining Multiple Runs	Mantyla et al. [69]	Clustering multiple outputs of the LDA model
	Belford et al. [11]	Combined multiple outputs of an NMF model
	Miller and McCoy [72]	Hierarchical clustering of the LDA model's output
	LTMA [32]	Combining topics of multiple runs using a hierarchical scheme
Initialization	STM [100]	Used LDA for initialization
	Belford et al. [11]	Used NNDSVD to initialize an NMF model
Parameter Adjustment	LDADE [1]	Used DE to obtain best settings

Table 2. List of techniques focused on improving stability. Each technique focused on at most two aspects to improve stability, as listed under the Type column. A short description of the technique employed to improve stability is provided in the right column.

7 DISCUSSION AND FUTURE DIRECTIONS

Stability, while important, is not the only desirable property of a topic model. To complement this paper's focus on metrics for measuring stability and techniques for improving stability, this section discusses other important factors that researchers can and perhaps should consider alongside measuring or improving stability. Specifically, this section addresses various impacts that corpus-level attributes can have on stability, the relationships between stability and other measures applied to topic models (e.g., quality, uniqueness and/or generality), and how stability can be applied to inform the use of topic modeling across various applications.

7.1 Corpus Effect

One source of instability in topic modeling can be the documents of a corpus. Different document orders generate different topic solutions as discussed in the previous Sections 1 and 6.4. Other issues related to documents and corpora can affect topic modeling performance as well. For example, Chakrabarti et al. [21] shows rare documents that does not share similar terms with other documents affect the performance of text mining models. Rare documents, or topics that are changing more often should be carefully determined and analyzed. Additionally, pre-processing methdos (e.g., corpus and token curation) or post-processing methods (e.g., author-topic entropy) can also affect the resultant topics [104, 119] and perhaps the stability of the models.

Alternatively, the relative lengths of each document (e.g., documents with varied lengths or documents with relatively short length) may affect the stability of topic modeling. For example, in dealing with long texts (e.g., novels), one can divide paragraphs, pages, chapters, or fixed number of tokens into separate documents [102, 126]. Any of these different approaches may affect topic solutions and the stability of a model. Furthermore, corpora with short length documents require extra attention. Word co-occurrence in short texts brings little additional information in topic modeling and can cause sparsity [64, 93, 131]. Moreover, labeling (inferencing) new incoming documents on the scale of, e.g., social media data with high variation of word co-occurrences is a critical issue [24]. In analyzing stability, no single work has been done to analyze the effects of length of documents on topic stability.

7.2 Comparing Stability with Other Measures

Stability should be seen as a property distinct from other measures, such as generality [85], uniqueness [31], or quality (e.g., coherence [22] or NPMI [61]). A highly stable model may or may not maintain high quality (i.e., more comprehensible to human users), low generality topics. Thus, other metrics should also be measured and interpret jointly.

For example, With high generality in topic modeling, multiple topics of a given model will have many top terms in common [85]. A topic model that yields high generality will also have stability, not because the same topics recur across multiple runs, but because the topics the model produces are all self-similar, even with the same run. Thus, maximizing only stability may result in topics that are so general as to be uninformative.

It is also important to consider how stability relates with topic quality. Improving stability does not necessarily improve the quality of the generated topics and vice versa. Is it preferable for a model to produce similar or even identical topics across multiple runs if those topics have poorer quality? How might different uses or applications of topic model [18] suggest making different tradeoffs between quality and stability. Furthermore, topic quality can be assessed using different methods, such as Perplexity [96], Coherence [78], Normalized Point-wise Mutual Information (NPMI) [75], and word-embedding evaluation methods [35], among others †. Each of these quality metrics and stability may function differently and should not be used as the only metric to compare performance of topic modeling techniques.

This subsection considers how generality [55, 85] and quality [61] may relate with stability. Future work, though, should examine closely the relationships between stability and a variety of other measures.It may be preferable to jointly optimize multiple metrics – for instance, high quality, high stability, and low generality – even if the result means a less-than-optimal value for a given one of these metrics. Indeed, combining some of these different metrics can provide deeper knowledge and clearer interpretation for comparing models, as described below.

7.3 Interpreting and Tuning Topic Models with Stability

This section reviews recent works on comparing and/or combining stability with other measures and discusses the need for exploring more in utilizing stability with other assessment methods.

As on example of comparing stability with quality, De Waal and Barnard [28] compared Perplexity (quality) to document correlation (stability) over different sizes of vocabulary of a corpus. They divided the corpus into training and test sets (80%-20%) and reduced the vocabulary size from 100% to 30%. Authors executed two GaP topic models on the trained data set and inferred topics for documents in the test set. Later, authors computed stability and quality of the inferred topics and it has been shown that the topic stability reduced less drastically than topic quality as the vocabulary size decreases. Till now, other than this work on comparing changes of stability and quality across different vocabulary size, there is no single work on examining the performance of quality and stability measures at the same time .

Very few works, only one to the best of authors' knowledge, introduced a topic modeling technique to improve both stability and quality at the same time. Koltcov et al. [53] showed gLDA improved both Coherence score and Jaccard stability in comparison with pLSA, LDA, and SLDA. However, little exploration has been done in comparing quality and stability regarding to different corpora, quality, or stability metrics.

While combining stability and quality can offer deeper understanding of the compared models, it may also add more uncertainty to the analysis. Mimno et al. [75] showed that the Point-wise Mutual Information (PMI) evaluation method is not stable across multiple runs. One way to resolve this is via averaging, either a single quality measure across multiple runs, or multiple quality measures across a single or multiple runs. However, Xing et al. [127] showed that the instability of the average of multiple metrics (including quality and stability metrics) gets higher with higher quality topics assessed by human subjects. These two papers are clear examples that utilizing stability with other metrics has no single solution and should be interpret based on the specific application for which topic modeling is used for.

†Besides these measures, there exists other evaluation methods to compute how a model is a good fit for the unseen data distribution (e.g., log-likelihood, or posterior checks[73]). But, because that is different than assessing a topic's quality itself, it was not brought into this comparison.

A set of guidelines to apply, use, and read topic modeling metrics may be needed with these caveats and uncertainty of relationships between quality and stability or even different interpretation of stability changes and values. Such guidelines may not be as straight forward as other machine learning performance assessment techniques.

Alternatively, human assessment is an alternative approach to compare quality of different machine learning models [61, 62, 101]. A similar approach can be employed to compute and compare stability [26]. At the time of writing, there is only one example of work that combined human assessment of stability and quality within one single run on one single model. Chuang et al. [26] introduced TopicCheck that allows for seeing relationships between computable metrics (e.g., quality) and human assessments of or changes to topics. This way users can see which topics are more stable based on document-topic distribution, and which ones are more readable (assuming that higher readability indicating higher quality) based on their top terms. As a future work, human perception of quality and stability can be compared to see if they function similarly or differently. This would be beneficial as topic modeling is meant to improve human readability in high number of applications.

As discussed earlier in this section, interpreting stability and utilizing it should be done based on purpose of a study and its application. As an example, for visualization and quantifying (e.g., predicting an output using document-topic distribution) applications stability may be more important than quality. It becomes difficult to find effective topics as factors in predicting an output with instable solutions. In contrast, for exploration purposes (e.g., finding trends and topics that were not found before), quality becomes more important than stability. Finding higher quality topics in each run that are not seen in previous runs can help researchers explore further. Topics with higher stability in such a empricial exploration may become less desirable in contrast with the importance of finding topics with higher coverage. This means, topic modeling outputs and measured metrics need cautious and further investigation.

7.4 Limitations

This work has three primary limitations, which provides important directions for future work. First, choosing a comprehensive set of keywords and searching different corpora regarding these terms was challenging. The term *stability* is widely used in the identified papers to refer to similarity of topic modeling outputs; however, some of the papers included in our final corpus refer to the concept of stability using other terms; e.g., *variability* or *robustness*. Although we updated the keywords through our search process, including Identification and Screening processes (Figure 4), future work could explore using other keywords to identify articles that may have been overlooked in this review.

Second, this paper focuses on specific databases where technical topic modeling papers often appear. As a result, it may omit papers about topic modeling applications, which are published in a wide variety of venues across disciplines, from medicine [cf. 29, 47], to humanities [cf. 18, 50], to political sciences [cf. 98, 100]. While it is possible that our exclusive focus on technical venues may affect the coverage of this review, it also seems likely that the technical advances on which this review focuses (i.e., measures for stability, and techniques for improving stability) are more likely to be published in technical computing venues than in venues focused on other disciplines.

Third, this paper reviews and focuses on English languages papers. Doing so might have eliminated some of the eligible works related to topic stability. Future work is encouraged to investigate papers published in other languages to review other measures or techniques regarding topic stability and stability improvement.

8 CONCLUSION

The issue of stability in topic modeling affects the ability to generate similar topics across multiple runs with different random initializations. Although multiple techniques have been introduced to improve stability, there is

no single model that can resolve this issue completely. To provide a thorough understanding about the issue of stability, this paper surveys sources of instability, approaches to measure stability, and techniques intended to improve stability. Most prior work comparing the stability of different models has been done on a single, general purpose corpus, such as news documents. It is still not clear if stability functions similarly on different types of corpora (e.g., news documents vs. fictional documents) or across different length of documents (e.g., novels vs. tweets). Furthermore, with more than a dozen of stability metrics, there is no single agreed-upon metric for stability due to the varied functionality of each metric and the varied information each offers. Thus, future work should compare different stability metrics to understand their relationships more thoroughly.

While having high stability is a must in some application and is somewhat important in others, measuring stability alone is not the best way to assess a topic model. Stability must be assessed alongside quality [61], uniqueness or generality [85], and other properties to create a fuller understanding about the performance of a model. Furthermore, just as prior work has compared quality metrics against human perceptions of topic quality [46, 60–62, 101], future work would also benefit from comparing stability metrics against human perceptions of topic stability. Doing so may help not only determine which stability metrics are most useful in various settings, but it can also help ensure that stability is an informative property in terms of the application for which a topic model is being used.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant #IIS-1814909. The authors would like to thank Brian D. Davison, Dominic DiFranzo, and the anonymous reviewers for stimulating discussions and helpful suggestions.

REFERENCES

[1] Amritanshu Agrawal, Wei Fu, and Tim Menzies. 2018. What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology* 98 (2018), 74–88.

[2] Edoardo M Airoldi and Jonathan M Bischof. 2016. Improving and evaluating topic models and other models of text. *J. Amer. Statist. Assoc.* 111, 516 (2016), 1381–1403.

[3] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6, 1 (2015).

[4] Asim Ansari, Yang Li, and Jonathan Z Zhang. 2018. Probabilistic topic model for hybrid recommender systems: A stochastic variational Bayesian approach. *Marketing Science* 37, 6 (2018), 987–1008.

[5] Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics* 6 (2018), 107–119.

[6] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*. 280–288.

[7] Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models–going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, 1–10.

[8] Eric PS Baumer, Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul, and Micki McGee. 2020. Topicalizer: reframing core concepts in machine learning visualization by co-designing for interpretivist scholarship. *Human–Computer Interaction* (2020), 1–29.

[9] Eric P. S. Baumer, Shion Guha, Emily Quan, David Mimno, and Geri K Gay. 2015. Missing photos, suffering withdrawal, or finding freedom? How experiences of social media non-use influence the likelihood of reversion. *Social Media+ Society* 1, 2 (2015), 2056305115614851.

[10] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.

[11] Mark Belford, Brian Mac Namee, and Derek Greene. 2018. Stability of topic modeling via matrix factorization. *Expert Systems with Applications* 91 (2018), 159–169.

[12] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[13] David M Blei, John D Lafferty, et al. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 1, 1 (2007), 17–35.

[14] David M Blei and Pedro J Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 343–348.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[16] Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *Journal of machine learning research* 2, Mar (2002), 499–526.
[17] Christos Boutsidis and Efstratios Gallopoulos. 2008. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern recognition* 41, 4 (2008), 1350–1362.
[18] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval* 11, 2-3 (2017), 143–296.
[19] Ana Catarina Calheiros, Sérgio Moro, and Paulo Rita. 2017. Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management* 26, 7 (2017), 675–693.
[20] John Canny. 2004. GaP: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 122–129.
[21] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. 1998. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB journal* 7, 3 (1998), 163–178.
[22] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems (NIPS)*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 288–296.
[23] Chao Chen, Alina Zare, Huy N Trinh, Gbenga O Omotara, James Tory Cobb, and Timotius A Lagaunne. 2017. Partial membership latent Dirichlet allocation for soft image segmentation. *IEEE Transactions on Image Processing* 26, 12 (2017), 5590–5602.
[24] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 2928–2941.
[25] Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*. 612–620.
[26] Jason Chuang, Margaret E Roberts, Brandon M Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 175–184.
[27] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. 2010. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of computer science in China* 4, 2 (2010), 280–301.
[28] Alta De Waal and Etienne Barnard. 2008. Evaluating topic models with stability. In *Proceedings of the Symposium of the Pattern Recognition Association of South Africa (PRASA)*. Cape Town, South Africa, 79–84.
[29] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391–407.
[30] Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology* 38, 1 (2004), 188–230.
[31] Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. 2018. Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 374–384.
[32] Mennatallah El-Assady, Fabian Sperrle, Rita Sevastjanova, Michael Sedlmair, and Daniel Keim. 2018. LTMA: Layered topic matching for the comparative exploration, evaluation, and refinement of topic modeling results. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*. IEEE, 1–10.
[33] Abdelmoula El-Hamdouchi and Peter Willett. 1989. Comparison of hierarchic agglomerative clustering methods for document retrieval. *Comput. J.* 32, 3 (1989), 220–227.
[34] Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. 2005. Stability of randomized learning algorithms. *Journal of Machine Learning Research* 6, Jan (2005), 55–79.
[35] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using word embedding to evaluate the coherence of topics from Twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 1057–1060.
[36] James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. 2013. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 446–454.
[37] Eric W Fox, Ryan A Hill, Scott G Leibowitz, Anthony R Olsen, Darren J Thornbrugh, and Marc H Weber. 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental monitoring and assessment* 189, 7 (2017), 1–20.
[38] András Frank. 2005. On Kuhn's Hungarian method—a tribute from Hungary. *Naval Research Logistics (NRL)* 52, 1 (2005), 2–5.
[39] Alan E Gelfand. 2000. Gibbs sampling. *Journal of the American statistical Association* 95, 452 (2000), 1300–1304.
[40] Maria J Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal* 26, 2 (2009), 91–108.

[41] Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 498–513.
[42] David Hall, Dan Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 363–371.
[43] Moritz Hardt, Ben Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*. PMLR, 1225–1234.
[44] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 50–57.
[45] Amin Hosseiny Marani, Joshua Levine, and Eric PS Baumer. 2022. One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 768–779.
[46] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in Neural Information Processing Systems* 34 (2021), 2018–2033.
[47] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.
[48] Tianyan Jiang, Jian Li, Yuanbing Zheng, and Caixin Sun. 2011. Improved bagging algorithm for pattern recognition in UHF signals of partial discharges. *Energies* 4, 7 (2011), 1087–1101.
[49] Michael Kearns and Dana Ron. 1999. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation* 11, 6 (1999), 1427–1453.
[50] Sergei Koltcov. 2018. Application of Rényi and Tsallis entropies to topic modeling optimization. *Physica A: Statistical Mechanics and its Applications* 512 (2018), 1192–1204.
[51] Sergei Koltcov, Vera Ignatenko, and Olessia Koltsova. 2019. Estimating Topic Modeling Performance with Sharma–Mittal Entropy. *Entropy* 21, 7 (2019), 660.
[52] Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. 2014. Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science*. 161–165.
[53] Sergei Koltcov, Sergey I Nikolenko, Olessia Koltsova, and Svetlana Bodrunova. 2016. Stable topic modeling for web science: granulated LDA. In *Proceedings of the 8th ACM Conference on Web Science*. 342–343.
[54] Olessia Koltsova and Sergei Koltcov. 2013. Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet* 5, 2 (2013), 207–227.
[55] Katsiaryna Krasnashchok and Aymen Cherif. 2019. Coherence regularization for neural topic models. In *International Symposium on Neural Networks*. Springer, 426–433.
[56] Fedor Krasnov and Anastasiia Sen. 2019. The number of topics optimization: clustering approach. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 416–426.
[57] Ludmila I Kuncheva and Dmitry P Vetrov. 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE transactions on pattern analysis and machine intelligence* 28, 11 (2006), 1798–1808.
[58] S Kutin and P Niyogi. 2002. *Almost-everywhere algorithmic stability and generalization error: Tech. Rep.* Technical Report. TR-2002-03: University of Chicago, Computer Science Department.
[59] Jey Han Lau and Timothy Baldwin. 2016. The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics, San Diego, California, 483–487.
[60] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 1536–1545.
[61] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.
[62] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017), 28–42.
[63] Shengqiao Li, E James Harner, and Donald A Adjeroh. 2011. Random KNN feature selection-a fast and stable alternative to Random Forests. *BMC bioinformatics* 12, 1 (2011), 450.
[64] Ximing Li, Yue Wang, Ang Zhang, Changchun Li, Jinjin Chi, and Jihong Ouyang. 2018. Filtering out the noise in short text topic modeling. *Information Sciences* 456 (2018), 83–96.
[65] Nathan C Lindstedt. 2019. Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. *Social Currents* 6, 4 (2019), 307–318.
[66] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5, 1 (2016), 1–22.

[67] Lars Maaloe, Morten Arngren, and Ole Winther. 2015. Deep belief nets for topic modeling. *arXiv preprint arXiv:1501.04325* (2015).
[68] Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. 2018. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures* 12, 2-3 (2018), 93–118.
[69] Mika V Mantyla, Maelick Claes, and Umar Farooq. 2018. Measuring LDA topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. 1–4.
[70] Vineet Mehta, Rajmonda S Caceres, and Kevin M Carter. 2014. Evaluating topic quality using model clustering. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 178–185.
[71] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 490–499.
[72] John Miller and Kathleen F McCoy. 2017. Topic model stability for hierarchical summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 64–73.
[73] David Mimno and David Blei. 2011. Bayesian checking for topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 227–237.
[74] David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 411–418.
[75] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.
[76] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and the PRISMA Group*. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* 151, 4 (2009), 264–269.
[77] Christopher E Moody. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019* (2016).
[78] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 100–108.
[79] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*. 215–224.
[80] Jiazhong Nie, Runxin Li, Dingsheng Luo, and Xihong Wu. 2007. Refine bigram PLSA model by assigning latent topics unevenly. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 141–146.
[81] Andreas Niekler and Patrick Jähnichen. 2012. Matching results of latent dirichlet allocation for text. In *Proceedings of ICCM*. 317–322.
[82] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science* 43, 1 (2017), 88–102.
[83] John D. O'Brien, Kathryn Lin, and Scott MacEachern. 2016. Mixture Model of Pottery Decorations from Lake Chad Basin Archaeological Sites Reveals Ancient Segregation Patterns. *Proceedings of the Royal Society B: Biological Sciences* 283, 1827 (March 2016), 20152824. https://doi.org/10.1098/rspb.2015.2824
[84] Makbule Ozsoy, Ilyas Cicekli, and Ferda Alpaslan. 2010. Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*. 869–876.
[85] Derek O'callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* 42, 13 (2015), 5645–5657.
[86] Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 2 (1994), 111–126.
[87] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery* 88 (2021), 105906.
[88] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. 2014. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 2 (2014), 256–270.
[89] Hae-Sang Park and Chi-Hyuck Jun. 2009. A simple and fast algorithm for K-medoids clustering. *Expert systems with applications* 36, 2 (2009), 3336–3341.
[90] Miha Pavlinek and Vili Podgorelec. 2017. Text classification method based on self-training and LDA topic models. *Expert Systems with Applications* 80 (2017), 83–93.
[91] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
[92] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 569–577.

[93] Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. 2017. Topic modeling over short texts by incorporating word embeddings. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 363–374.
[94] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering* (2020).
[95] Ueli Reber. 2019. Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication methods and measures* 13, 2 (2019), 102–125.
[96] Loïs Rigouste, Olivier Cappé, and François Yvon. 2007. Inference and evaluation of the multinomial mixture model for text clustering. *Information processing & management* 43, 5 (2007), 1260–1280.
[97] Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. 2016. A model of text for experimentation in the social sciences. *J. Amer. Statist. Assoc.* 111, 515 (2016), 988–1003.
[98] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2016. Navigating the local modes of big data. *Computational Social Science* 51 (2016).
[99] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An R package for structural topic models. *Journal of Statistical Software* 91, 1 (2019), 1–40.
[100] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58, 4 (2014), 1064–1082.
[101] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.
[102] Stefano Sbalchiero and Maciej Eder. 2020. Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Quality & Quantity* (2020), 1–14.
[103] Simone Scardapane and Dianhui Wang. 2017. Randomness in neural networks: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7, 2 (2017), e1200.
[104] Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 432–436.
[105] Erich Schubert and Peter J Rousseeuw. 2019. Faster k-Medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In *International Conference on Similarity Search and Applications*. Springer, 171–187.
[106] Fanhua Shang, Zhihui Zhang, Yuying An, Yang Hu, and Hongying Liu. 2019. Efficient Parallel Stochastic Variance Reduction Algorithms for Large-Scale SVD. In *2019 International Conference on Data Mining Workshops (ICDMW)*. 172–179. https://doi.org/10.1109/ICDMW.2019.00035
[107] Deepak Sharma, Bijendra Kumar, and Satish Chand. 2017. A survey on journey of topic modeling techniques from SVD to deep learning. *International Journal of Modern Education and Computer Science* 9, 7 (2017), 50.
[108] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 63–70.
[109] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics* 5 (2017), 1–16.
[110] David Sontag and Dan Roy. 2011. Complexity of inference in latent dirichlet allocation. *Advances in neural information processing systems* 24 (2011).
[111] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961.
[112] Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis* 427, 7 (2007), 424–440.
[113] Rainer Storn. 1996. On the usage of differential evolution for function optimization. In *Proceedings of North American Fuzzy Information Processing*. IEEE, 519–523.
[114] Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* 3, Dec (2002), 583–617.
[115] Wei Sun. 2015. *Stability of machine learning algorithms*. Ph.D. Dissertation. Purdue University.
[116] Matt Taddy. 2012. On estimation and selection for topic models. In *Artificial Intelligence and Statistics*. 1184–1193.
[117] Jie Tang, Ruoming Jin, and Jing Zhang. 2008. A topic modeling approach and its integration into the random walk framework for academic search. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 1055–1060.
[118] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*. 1385–1392.

[119] Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3903–3914.
[120] Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*. 1973–1981.
[121] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*. 1105–1112.
[122] Yinying Wang, Alex J Bowers, and David J Fikis. 2017. Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational administration quarterly* 53, 2 (2017), 289–323.
[123] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.
[124] Lino Wehrheim. 2019. Economic history goes digital: topic modeling the Journal of Economic History. *Cliometrica* 13, 1 (2019), 83–125.
[125] Ryan Wesslen. 2018. Computer-assisted text analysis for social science: Topic models and beyond. *arXiv preprint arXiv:1803.11045* (2018).
[126] Zongda Wu, Li Lei, Guiling Li, Hui Huang, Chengren Zheng, Enhong Chen, and Guandong Xu. 2017. A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications* 84 (2017), 12–23.
[127] Linzi Xing, Michael J Paul, and Giuseppe Carenini. 2019. Evaluating Topic Quality with Posterior Variability. *arXiv preprint arXiv:1909.03524* (2019).
[128] Yi Yang, Shimei Pan, Jie Lu, Mercan Topkara, and Yangqiu Song. 2016. The stability and usability of statistical topic models. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 2 (2016), 1–23.
[129] Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 937–946.
[130] Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of Biostatistics* 7 (2005).
[131] Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2105–2114.