Journal of Machine Learning Research 23 (2022) 1-20 Submitted 4/19; Revised 3/21; Published 2/22

A Statistical Approach for Optimal Topic Model Identification

Craig M. Lewis CRAIG.LEWIS@OWEN.VANDERBILT.EDU
Owen Graduate School of Management
Vanderbilt University, Nashville, TN, USA

Francesco Grossetti FRANCESCO.GROSSETTI@UNIBOCCONI.IT
Department of Accounting and
Bocconi Institute for Data Science and Business Analytics (BIDSA)
Bocconi University, Milan, Italy

Editor: David Blei

Abstract

Latent Dirichlet Allocation (LDA) is a popular machine-learning technique that identifies latent structures in a corpus of documents. This paper addresses the ongoing concern that formal procedures for determining the optimal LDA configuration do not exist by introducing a set of parametric tests that rely on the assumed multinomial distribution specification underlying the original LDA model. Our methodology defines a set of rigorous statistical procedures that identify and evaluate the optimal topic model. The U.S. Presidential Inaugural Address Corpus is used as a case study to show the numerical results. We find that 92 topics best describe the corpus. We further validate the method through a simulation study confirming the superiority of our approach compared to other standard heuristic metrics like the perplexity index.

Keywords: topic modeling, Latent Dirichlet Allocation, model selection, parametric testing, optimization

1. Introduction

Textual analysis has been widely used in a number of different contexts across a wide range of disciplines that include, for example, finance, accounting, marketing, health care, and, even, movie choices (Rubin and Syeyvers, 2006; Core et al., 2008; Larcker and Zakolyukina, 2012; Lu et al., 2016; Toubia et al., 2019). A category of existing approaches is known as *Bag-of-Words* (BoW) techniques and typically rely on simple word counts rather than evaluating word choices in their intended context. This implies that the order of the words is not necessary to describe a document. Textual analysis is a broad discipline and its techniques vary in terms of scope and complexity. Examples include:

- Calculating the number of words that are contained in topic-specific dictionaries; Examples of recent work include Li (2006); Core et al. (2008); Loughran et al. (2009); Larcker and Zakolyukina (2012);

- Metrics that attempt to discern the clarity of text such as the Fog Index (Gunning, 1969) and the Flesch-Kincaid score (Kincaid et al., 1975). These types of measures are also called readability indexes;

- Supervised learning approaches like Naïve Bayes classification (Russell and Norvig, 2016; Rish et al., 2001; Li, 2010);

- Unsupervised topic modeling techniques such as *Latent Dirichlet Allocation* (LDA, (Blei et al., 2003));

- Word embedding techniques that capture a large number of syntactic and semantic word relationships by building a vector representation of the corpus (Mikolov et al., 2013).

The focus of this paper is topic modeling, a statistical technique that allows the researcher to extract latent features, called *topics*, from a collection of textual documents. One of the most important advantages of topic modeling is its inherent statistical nature. The identified topics and their atomic components, called *words*, are both drawn from probability distributions.

The genesis for this family of models began with the development of Latent Semantic Indexing (LSI) by Deerwester et al. (1990). This approach uses a singular value decomposition to extract uncorrelated topics much like principal components analysis. Hofmann (1999, 2017) extends LSI by specifying a generative model for the data that treats topics as probability distributions over words. This intuition enables the researcher to disentangle polysemous words so that one can recognize their potential different meanings. Blei et al. (2003) further extend this framework by developing a Bayesian version of the probabilistic LSI (pLSI) model, called Latent Dirichlet Allocation (LDA), to also include a statistical model at the level of documents.[1]

LDA is a generative statistical model that identifies narrative topics from a corpus of documents under the assumption that the document-topic and topic-word distributions have Dirichlet priors. It relies on the intuition that a document can be represented by a set of common topics and that the content of a specific document can be described by the weights that are placed on these topics. The generative nature of LDA is a key advantage because it does not require researcher pre-judgment and is replicable. In this sense, it differs from dictionary-based approaches that rely on *ad-hoc* lists of words that are developed by researchers to represent pre-specified thematic content. For example, Loughran and McDonald (2011) develop a number of finance-specific dictionaries that classify words according to their narrative content. They include lists of positive words, negative words, uncertainty words, litigious words, strong and weak modal words.[2]

A limitation of LDA, as well as of any other unsupervised method, is that the optimal number of topics is unknown a-priori. As Gerlach et al. (2018) note:

1. LDA assumes that topics are uncorrelated. Although it is not a focus of this paper, Blei et al. (2007) extend LDA by introducing a specification that accommodates correlated topics.
2. The general word lists can be found here: `https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists`.

> *[...]Despite its success and overwhelming popularity, LDA is known to suffer from fundamental flaws in the way it represents text. In particular, it lacks an intrinsic methodology to choose the number of topics and contains a large number of free parameters that can cause overfitting.[...]*

This is problematic because different specifications will likely lead to different interpretations of the corpus. Since researchers using LDA must pre-specify the number of topics to be estimated, an underspecified model is too coarse to be useful in uncovering the underlying structure, while a model that estimates too many topics could instead generate uninformative and possibly redundant topics.

In their work, Gerlach et al. (2018) overcome these issues by relying on a different approach. They represent the word-document matrix as a bipartite network which makes the problem of estimating the topics equivalent to finding communities. They then develop a formal correspondence that builds on the mathematical equivalence of pLSI and Stochastic Block Models (SBMs) (Holland et al., 1983; Airoldi et al., 2008; Ball et al., 2011; Karrer and Newman, 2011). In contrast to Gerlach et al. (2018), our paper relies on the classic LDA framework as described in Blei et al. (2003). We define a parametric test that builds and exploits the mathematical constructs as defined in the original LDA setting. Rather than modify the estimation procedure, as in Gerlach et al. (2018), we develop a simple parametric approach that identifies the optimal topic specification *ex-post*. This provides the researcher with an internally consistent statistical framework for optimal topic selection that only requires LDA estimates.

In the context of topic modeling, there exist several *ad-hoc* evaluation strategies that provide guidance on how to identify the "optimal" number of topics. For instance, it is common for researchers to run a series of models with slightly different specifications or use cross-validation on hold out document sets. For example, Zhao et al. (2015) describe how iterative approaches can evaluate alternative specifications using cross-validation on hold out data. Two additional relatively common approaches were introduced by Cao et al. (2009) and Arun et al. (2010) and consist of a density-based clustering and a matrix factorization exploiting KL-divergence, respectively. Another standard, and far more intuitive, approach determines which specification is the least *perplexed* by the test sets. Perplexity is based on the intuition that a high degree of similarity, identified as a low level of perplexity, can be used to determine the appropriate number of topics (Blei et al., 2003; Hornik and Grün, 2011). Formally, for a test set of J documents, perplexity is defined as:

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{j=1}^{J}\log\left[Pr(D_j)\right]}{\sum_{j=1}^{J}P_j}\right\}, \tag{1}$$

where $Pr(D_j)$ is the probability of the observed document D_j and P_j is the number of words in document D_j.

This paper develops a formal parametric solution to optimal topic identification that is rigorous yet intuitive. It is based on the conjecture that, if there exists a finite number of topics that fully characterize a corpus, they should explain the corpus better than any other specification and that additional topics should not provide significant incremental explanatory power. Under the assumption that the *optimal* topic distribution is optimally characterized by a set of K topics, it should display three properties:

Property 1. A K-topic model should fully characterize the word distributions for the documents in the corpus.

Property 2. The K most similar topics from a $\hat{K}$-topic model should be statistically indistinguishable from the topics estimated from the optimal K-topic model where $\hat{K} > K$. The K most similar topics from a $\hat{K}$-topic model are the K topics that have the highest cosine similarities with the K topics from the optimal topic model.

Property 3. The $\hat{K} - K$ topics that are least similar should not provide significant incremental explanatory power relative to the K most similar topics when compared to the actual word distributions for the documents in the corpus. The $\hat{K} - K$ least similar topics are those that are not identified as being most similar.

More recent works like Gerlach et al. (2018) and Fortunato (2010) develop a nonparametric approach for determining the number of topics by exploiting the parallelism between topic models and community detection methods. In particular, Gerlach et al. (2018) argue that the assumption of Dirichlet priors is a conceptual limitation of LDA models as it is not always supported by the data and the inability to identify the number of topics is a practical limitation.

Our parametric approach addresses this second concern. It relies on the intuition that, as one increases the number of estimated topics, a point is reached where K topics are sufficient to explain the narrative content of the corpus. We call this concept *topic stability*.

The tests we describe are rigorous and follow directly from the classic LDA specification as given in Blei et al. (2003). In this sense, our methodology is fully parametric. It also is heuristic because it appeals to intuitive concepts that should be expected to hold if the model conforms to our conjecture about topic stability.

2. LDA: the Basic Setup

LDA (Blei et al., 2003) is based on the idea that a corpus can be represented by a set of topics. LDA uses a likelihood approach to discover latent clusters of text, namely *topics*, that frequently appear in a corpus. The method assumes that the document generation process arises from an underlying topic distribution rather than a distribution over individual words. A particular topic can be characterized as a distribution over a common vocabulary of words where the relative probability weight assigned to each word indicates its relative importance to that topic.

We refer to the probability weights assigned to specific words as Topic Word Weights (**TWW**s). A topic is thus a word vector where each element represents that word's relative importance to the topic. For example, the words "oil" and "electricity" might be important to topics associated with Natural Resources and Manufacturing, but one might expect oil to receive a higher weighting than electricity in the Natural Resources topic. The opposite might be true for the Manufacturing topic. Each document is then represented as a linear combination of different topics. We refer to the weights applied to each topic within a specific document as a vector of Document Topic Weights (**DTW**).

If one assumes that the words in different documents are drawn from K topics, the distribution of words can then be characterized as a mixture of these topics such that the probability of observing word w_p is:

$$Pr(w_p) = \sum_{k=1}^{K} Pr(w_p|z_p = k)\, Pr(z_p = k) \qquad \text{with } p = 1, \ldots, P, \tag{2}$$

where z_p is a latent variable that indicates the topic from which w_p was drawn. $Pr(w_p|z_p = k)$ is the probability of w_p in the k-th topic, and $Pr(z_p = k)$ is the probability that the word is drawn from that same k-th topic.[3]

The observable data are contained in a corpus denoted by $\mathbf{D}$ made of J documents such that $\mathbf{D} = \{D_1, \ldots, D_J\}$. Each document D_j is a sequence of P_j words such that $D_j = \{w_1, \ldots, w_{P_j}\}$. Conceptually, a document D_j is generated by drawing a topic k from the topic distribution and then word w_p from the word distribution conditional on topic k.[4]

The model is formalized by assuming that for each document D_j there is a multinomial distribution over the K topics with parameter vector θ_j^K. This implies that word w_p in document D_j is selected from topic k with probability $Pr(z_p = k) = \theta_{jk}^K$. Intuitively, when we aggregate this probability at the corpus level we obtain a $J \times K$ matrix θ^K which represents the **DTW**s. For each topic k there is a multinomial distribution over P words with parameter vector ϕ_p^K such that $Pr(w_p|z_p = k) = \phi_{pk}^K$. By collecting the parameter vectors, we obtain a $K \times P$ matrix $\boldsymbol{\phi}^K$ which represents the **TWW**s.

LDA estimation is conducted by choosing the optimal values of $\boldsymbol{\theta}^K$ and $\boldsymbol{\phi}^K$. To make predictions about the corpus $\mathbf{D}$, both $\boldsymbol{\theta}^K$ and $\boldsymbol{\phi}^K$ are assumed to have Dirichlet prior distributions with respective scalar hyper-parameters α and β. The Dirichlet distribution is a natural choice because it is the conjugate prior to the multinomial distribution.

Following Blei et al. (2003) and suppressing hyper-parameters α and β for expositional clarity, the generative process for LDA corresponds to the following joint distribution of latent and observed variables:

$$Pr(\mathbf{D}, \mathbf{Z}, \boldsymbol{\theta}^K, \boldsymbol{\phi}^K) = \prod_{j=1}^{J} Pr\left(\theta_j\right) \prod_{k=1}^{K} Pr\left(\phi_k^K\right) \left[\prod_{p=1}^{P} Pr\left(z_{jp}|\theta_j^K\right) Pr\left(w_{jp}|z_{jp}, \phi^K\right)\right], \tag{3}$$

where $\mathbf{Z}$ denotes a $P \times J$ matrix where each element z_{jp} is the topic assignment for the p-th word in document D_j.

The output of a LDA estimation is represented by a $J \times K$ **DTW** matrix and a $K \times P$ **TWW** matrix. Due to the coupling of $\boldsymbol{\theta}^K$ and $\boldsymbol{\phi}^K$, exact inference is intractable (Dickey, 1983). Various approximate algorithms such as variational inference or Markov Chain Monte Carlo are typically used for inference (Jordan, 1998). This paper uses the *Variational Expectation-Maximization* (VEM) method (Jordan et al., 1999).

3. For an excellent discussion of LDA, see the paper by Griffiths and Steyvers (2004).
4. This intuition forms the basis for estimating LDA models that rely on a Monte Carlo Markov Chain simulation coupled with a Gibbs sampler.

3. Identification of the Optimal K-Topic Model

Our initial test of overall model adequacy evaluates how well a K-topic model explains the corpus. To do this, we describe a chi-square test that identifies the optimal number of topics. The test relies on the observation that each word in document D_j can be represented as:

$$d_j = \sum_{k=1}^{K} \theta_{jk}^K \phi_k^K + \epsilon_j, \tag{4}$$

where d_j is a $1 \times P$ row vector of word proportions associated with document D_j and θ_{jk}^K is a $J \times 1$ column vector of **DTW**s associated with k-th topic and document D_j. The test statistic in Equation (4) is used by estimating different sized topic models and then selecting the topic structure that most closely matches the word proportions in the underlying corpus.

To test the adequacy of different specifications, we define a Pearson chi-square statistic (Agresti, 1996) that exploits the underlying assumption that d_j and $\sum_{k=1}^{K} \theta_{jk}^K \phi_k^K$ are distributed multinomial. If a K-topic model fully characterizes the corpus, we would be unable to reject the hypothesis that the observed and estimated word distributions for the document D_j are statistically indistinct.

Due to the large number of **TWW**s with near zero probabilities, we collapse relatively unimportant words into a single bin if $I_{jp}^K < I_{j,\min}$ where:[5]

$$I_{jp}^K = \sum_{k=1}^{K} \theta_{jpk}^K \phi_{pk}^K. \tag{5}$$

The probability of observing a relatively unimportant word related to document D_j is then defined as:

$$I_{j,\min}^K = \sum_{p \in \{I_{jp}^K < I_{j,\min}\}} I_{j,p}^K, \tag{6}$$

and the actual frequencies of observing the same set of words in document D_j is:

$$D_{j,\min} = \sum_{p \in \{I_{jp}^K < I_{j,\min}\}} D_{j,p}. \tag{7}$$

Note that, even though each document has a unique number of relatively unimportant words, the basis for identifying the cutoff value is the LDA model rather than the actual documents in the corpus. Also note that, by collapsing the relatively unimportant words in this manner we preserve the underlying assumption that **TWW**s are distributed multinomial. The chi-square statistic is then calculated as follows:

Test 1 (Aggregate DTW and TWW Stability) *A K-topic model fully characterizes the corpus if the observed and estimated word vectors are statistically indistinct. The test statistic for the corpus is:*

5. We identify $I_{j,\min}$ as the smallest value of I_{jp}^K such that $\sum_{p=1}^{\hat{p}} I_{jp}^K < I^K$, where I^K is a minimum probability cutoff. We use $I^K = 0.05$ in the numerical example.

$$\text{OpTop}_J^K = \sum_{j=1}^{J} \left[(P_j + 1) \left(\sum_{p=1}^{P_j} \frac{\left(D_{jp} - I_{jp}^K \right)^2}{I_{jp}^K} + \frac{\left(D_{j,\min} - I_{j,\min}^K \right)^2}{I_{j,\min}^K} \right) \right] \sim \chi^2_{P_J}, \quad (8)$$

where OpTop_J^K *is distributed chi-square with* $P_J = \sum_{j=1}^{J} P_j$ *degrees of freedom and* P_j *is the number of relatively important words in document* D_j*.*

4. Topic Stability

A corpus is said to be K-topic stable (*Property 2*) when the K topics that best characterize its narrative content do not change as additional topics are added. **TWW**-stability is a necessary but not sufficient condition for a corpus to be K-topic stable. It is defined as follows:

Definition 1 (TWW Stability) *The* ***TWW****s are deemed "stable" when the absolute difference between* ϕ^K *and the* K *most similar topics from a* $\hat{K}$*-topic model,* $\phi^{\hat{K}}(\kappa)$*, approaches zero where* $\hat{K} > K$[6]*. Formally we have:*

$$\sum_{k=1}^{K} \sum_{p=1}^{P} \left| \phi_{pk}^K - \phi_{pk}^{\hat{K}}(\kappa) \right| \to 0, \quad (9)$$

where $\phi^{\hat{K}}(\kappa)$ *is the subset of* $\hat{K}$ *topics that have the highest cosine similarity with* ϕ^K*, i.e.,:*

$$\phi_k^{\hat{K}}(\kappa) = \max_{p \in \{1,\ldots,\hat{K}\}} \phi_k^{K^T} \phi_p^{\hat{K}}, \quad \forall k \in \{1, \ldots, \hat{K}\}. \quad (10)$$

Since LDA is based on the assumption that each topic follows an independent multinomial distribution, we can test whether the **TWW** distributions for each of the individual k topics from a K-topic model are stable relative to the K most similar topics from a $\hat{K}$-topic model using a Pearson chi-square test.

Once again, we mitigate the influence of relatively unimportant words by collapsing all words that have **TWW**s less than $\phi_{\min}$ into a single bin. The probability of observing uninformative words related to topic k is:

$$\phi_{k,\min}^K = \sum_{p \in \{\phi_{pk}^K < \phi_{k,\min}\}} \phi_{p,k}^K, \quad (11)$$

and the corresponding frequency of observing uninformative words in the most similar topic from a $\hat{K}$-topic model is:

$$\phi_{k,\min}^{\hat{K}} = \sum_{p \in \{\phi_{pk}^K < \phi_{k,\min}\}} \phi_{p,k}^{\hat{K}}(\kappa). \quad (12)$$

6. Our definition of most similar topics is cosine similarity. This measure is commonly used to assess how close two vectors are to each other (Singhal et al., 2001).

Note that, the relatively unimportant words that comprise $\phi^{\hat{K}}_{k,\min}$ are identified relative to the "optimal" *K-topic* model.

Test 2 (*k*-th Topic TWW Stability) *The k-th topic from a K-topic LDA model is stable relative to its most similar topic from a $\hat{K}$-topic model if one cannot reject the hypothesis that TWW^K_k is statistically different from zero where:*

$$TWW^K_k = (P_k + 1)\left[\sum_{p=1}^{P_k} \frac{\left(\phi^{\hat{K}}_{pk}(\kappa) - \phi^K_{pk}\right)^2}{\phi^K_{pk}} + \frac{\left(\phi^{\hat{K}}_{k,\min}(\kappa) - \phi^K_{k,\min}\right)^2}{\phi^K_{k,\min}}\right] \sim \chi^2_{P_k}, \tag{13}$$

and P_k is the number of words in the vocabulary that have $\phi^K_{pk} \geq \phi_{\min}$. TWW^K_k is distributed chi-square with P_k degrees of freedom.

Since the sum of K chi-square distributions is a chi-square, we can formally test whether the corpus displays aggregate K-topic stability by summing across all K topics:[7]

Test 3 (Aggregate TWW Stability) *The K topics from a K-topic model are stable relative to their most similar topics from a $\hat{K}$-topic model if one cannot reject the hypothesis that $\mathbf{TWW^K}$ is statistically different from zero where:*

$$\mathbf{TWW^K} = \sum_{k=1}^{K} TWW^K_k \sim \chi^2_{P_K}. \tag{14}$$

$\mathbf{TWW^K}$ *is distributed chi-square with $P_K = \sum_{k=1}^{K} P_k$ degrees of freedom.*

5. Tests of Overall Model Adequacy

The above set of tests allow the researcher to identify the K-topic model that best describes the corpus (**Test 1**) and to infer whether the **TWW**s are stable (**Tests 2** and **3**). These three tests are used to determine whether successive LDA iterations uncover the same set of topics. An implication of a corpus that displays topic stability is that K topics are sufficient to fully characterize its narrative content.

To test whether a specific K-topic model characterizes the narrative content of a corpus (*Property 3*), we next describe two overall goodness-of-fit tests. Both tests are based on the idea that the fitted values from a $\hat{K}$-topic model can be decomposed into *informative* and *uninformative* components.

7. An alternative test can be designed that relies on the Central Limit Theorem. Since the chi-square distribution is a sum of independent random variables, $TWW^K_{\mathbf{k}}$ converges to a normal distribution for large P. The K most similar topics across successive LDA estimations are stable if one cannot reject the hypothesis that $\mathbf{Z}^{TWW}_K$ is statistically indistinct from zero, i.e.,

$$\mathbf{Z^{TWW}_K} = \sum_{k=1}^{K} \frac{TWW^K_k - P}{\sqrt{2\sum_{k=1}^{K} P_k}} \to \mathcal{N}(0,1).$$

5.1 Identification of Informative and Uninformative Components

All of the tests in this section decompose the fitted values in Equation (4) into their *informative* and *uninformative* components. The informative component is calculated by selecting the K topics from a $\hat{K}$-topic model that are most similar to those from a K-topic model. The uninformative component then represents that portion of the fitted values associated with the remaining $\hat{K} - K$ topics.

We therefore test whether we can reject the hypothesis that the fitted values from an optimal K-topic model and the K most similar topics from a $\hat{K}$-topic model follow the same distribution. To do this, we replace θ_{jk}^{K} in Equation (4) with $\hat{\theta}_{jk}^{\hat{K}}(K)$. This yields a $P \times 1$ word vector that corresponds to the P words in the corpus dictionary such that:

$$\hat{I}_{j}^{\hat{K}} = \left\{ \frac{\sum_{k=1}^{K} \hat{\theta}_{jpk}^{\hat{K}}(\kappa)\phi_{pk}^{K}}{\sum_{p=1}^{P}\sum_{k=1}^{K} \hat{\theta}_{jpk}^{\hat{K}}(\kappa)\phi_{pk}^{K}} \right\}_{p=1}^{P}. \tag{15}$$

The uninformative component is calculated as:

$$\hat{U}_{j}^{\hat{K}} = \left\{ \frac{\sum_{k=1}^{\hat{K}-K} \hat{\theta}_{jpk}^{\hat{K}}(\kappa)\phi_{pk}^{K}}{\sum_{p=1}^{P}\sum_{k=1}^{\hat{K}-K} \hat{\theta}_{jpk}^{\hat{K}}(\kappa)\phi_{pk}^{K}} \right\}_{p=1}^{P}. \tag{16}$$

where $\theta_{jk}^{\hat{K}}(\kappa)$ is the set of $\hat{K}-K$ **TWW**s that are least similar to the **TWW**s from a K-topic model. We also scale the informative and uninformative components in Equations (15) and (16) by $\sum_{p=1}^{P}\sum_{k=1}^{K} \hat{\theta}_{jpk}^{\hat{K}}(\kappa)\phi_{pk}^{K}$ and $\sum_{p=1}^{P}\sum_{k=1}^{K-\hat{K}} \hat{\theta}_{jpk}^{\hat{K}}(\kappa)\phi_{pk}^{K}$, respectively, so that $\hat{I}_{j}^{\hat{K}}$ and $\hat{U}_{j}^{\hat{K}}$ both sum to one and can be interpreted as conditional multinomial distributions.

5.2 Chi-square Tests of Relative Information Content

Once it is determined that K topics are sufficient and stable (**Tests 1**, **2** and **3**), the next step is to test whether the information contained in the informative component from a $\hat{K}$-topic model closely tracks that from the optimal K-topic model.

5.2.1 Aggregate K-topic Stability

This procedure is different from **Test 3** because it only considers the fitted values associated with the informative component of a K-topic model. The test statistic is defined as:

Test 4 (Aggregate K-Topic Stability) *A K topic model fully characterize a corpus of* **D** *made of J documents if one cannot reject the hypothesis that the K most similar topics from a $\hat{K}$-topic model are statistically distinct. The test statistic* $\text{AGG\_STAB}_{\hat{I}}^{K}$ *is specified as:*

$$\text{AGG\_STAB}_{\hat{I}}^{K} = \sum_{j=1}^{J}\left[(P_j+1)\left(\sum_{p=1}^{P_j}\frac{\left(\hat{I}_{jp}^{K} - I_{jp}^{K}\right)^2}{I_{jp}^{K}} + \frac{\left(\hat{I}_{j,\min}^{K} - I_{j,\min}^{K}\right)^2}{I_{j,\min}^{K}}\right)\right] \sim \chi_{P_J}^{2}, \tag{17}$$

where P_j is the number of relatively important words in document D_j that have $I_{jp}^K \geq I_{\min}$. $\text{AGG\_STAB}_{\hat{I}}^K$ is distributed chi-square with $P_J = \sum_{j=1}^J P_j$ degrees of freedom.

5.2.2 Relative Importance of Incremental Topics

We next evaluate whether the $\hat{K} - K$ additional (least similar) topics contain incremental information that can be relevant in explaining the corpus. The idea is to compare how well the K most similar topics describe the corpus relative to the remaining $\hat{K} - K$ least similar ones. We introduce an F-test that is essentially a horse-race between the informative and uninformative components.

Test 5 (Relative Importance of Incremental Topics) *K topics adequately characterize a corpus $\mathbf{D}$ made of J documents if the incremental information contained in the uninformative component $\hat{U}_{jp}^K$ relative to the informative one $\hat{I}_{jp}^K$ is statistically indistinct from zero. The test statistic F_K^{CORP} is defined as follows:*

$$F_K^{\text{CORP}} = \frac{\text{INFORM}_{\hat{I}}^K}{\text{UNINFORM}_{\hat{U}}^K} \sim F_{(P_J, P_J)}, \tag{18}$$

where F_K^{CORP} is distributed as $F_{(P_J,P_J)}$, $P_J = \sum_{j=1}^J P_j$ is the number of degrees of freedom, and P_j is the number of relatively important words in document D_j. The chi-square statistics $\text{INFORM}_{\hat{I}}^K$ and $\text{UNINFORM}_{\hat{U}}^K$ consider whether the distributions implied by the informative component $\hat{I}_j^K$ and the uninformative one $\hat{U}_j^K$ for document D_j are similar to the distribution of document D_j's observed word proportions. They are defined as follows:

$$\text{INFORM}_{\hat{I}}^K = \sum_{j=1}^J \left[(P_j + 1) \left(\sum_{p=1}^{P_j} \frac{\left(D_{jp} - \hat{I}_{jp}^{\hat{K}} \right)^2}{\hat{I}_{jp}^{\hat{K}}} + \frac{\left(D_{j,\min} - \hat{I}_{j,\min}^{\hat{K}} \right)^2}{\hat{I}_{j,\min}^{\hat{K}}} \right) \right] \sim \chi_{P_J}^2, \tag{19}$$

$$\text{UNINFORM}_{\hat{U}}^K = \sum_{j=1}^J \left[(P_j + 1) \left(\sum_{p=1}^{P_j} \frac{\left(D_{jp} - \hat{U}_{jp}^{\hat{K}} \right)^2}{\hat{U}_{jp}^{\hat{K}}} + \frac{\left(D_{j,\min} - \hat{U}_{j,\min}^{\hat{K}} \right)^2}{\hat{U}_{j,\min}^{\hat{K}}} \right) \right] \sim \chi_{P_J}^2. \tag{20}$$

6. Case Study: the U.S. Presidential Inaugural Address Corpus

We test our algorithm on the U.S. presidential inaugural address texts (Peters, 2018). The corpus contains 58 documents of US president's inaugural addresses starting with George Washington's first inaugural address in 1789. Table 1 reports that the mean number of sentences per speech is 86. On average, each speech is comprised of 2,332 Tokens and 805 of these words are distinct (Types). This implies that each word is used approximately 2.9 times. While some documents are quite lengthy, others are very short.

	Mean	Std. Dev.	Percentiles 1%	25%	50%	75%	99%
Types	805	324	198	556	773	988	1,634
Tokens	2,332	1,382	376	1,434	2,084	2,892	6,726
Sentences	86	47	14	44	88	118	202

Table 1: Summary statistics for the 58 inaugural speeches by all the U.S. Presidents.

6.1 The Optimal K-topic Model — Test 1

As we mentioned at the beginning of the paper, the usual method to assess the optimal number of topics given a set of independent LDA models is the perplexity index as given in Equation (1). Figure 1 reports the in-sample perplexity index for the LDA models ranging from 2 to 200 topics which concludes that the best fit is given by 86 topics.

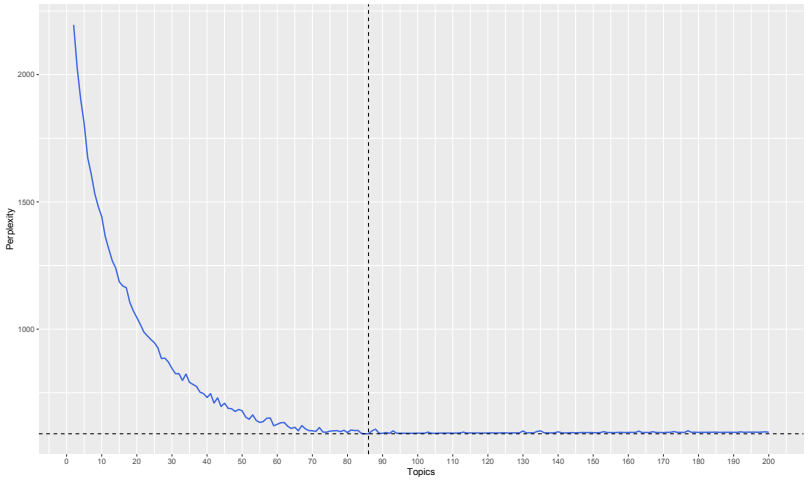

Figure 1: Perplexity index for LDA models ranging from 2 to 200 topics.

Test 1 introduced in Equation (8) formally identifies the topic model that best characterizes the corpus using a parametric test. We first consider how well different topic specifications explain individual documents. We then aggregate these results to consider the entire corpus. The overall conclusion is that an LDA model with 92 topics fits best. The optimum is identified as the K-topic model with the minimum $OpTop_J^K$ across all models ranging from 2 to 200 topics. Figure 2 shows the results for this initial test. Notice that the standardized $OpTop_J^K$ chi-square statistic tracks the perplexity measure but has the advantage of being a parametric test that assesses the goodness-of-fit.

6.2 Addressing Topic Stability — Tests 2 and 3

To provide a comprehensive picture of individual topic stability, Figure 3 illustrates the TWW_k^K statistics associated with **Test 2** for higher dimensional models that range from 93 to 200 topics relative to the identified optimal 92-topic model. Intuitively, we are looking for a substantially flat plot. As we can see from Figure 3, the vast majority is indeed flat

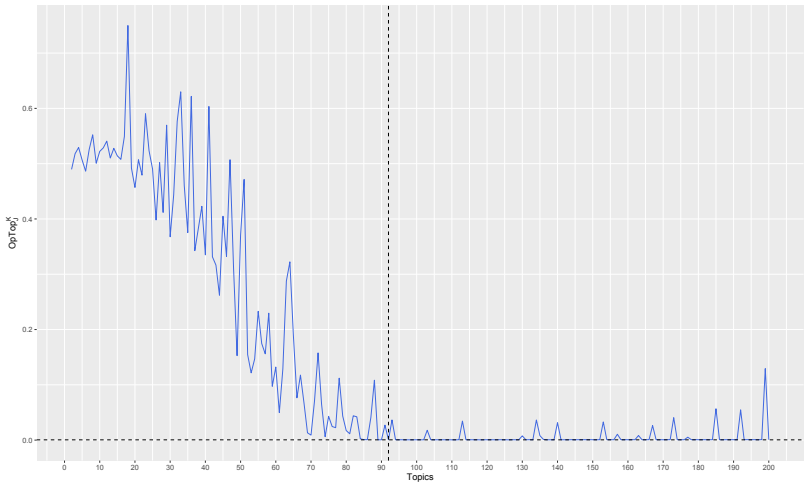

Figure 2: Standardized $OpTop_J^K$ statistics for **Aggregate DTW-TWW Stability** for LDA models ranging from 2 to 200 topics. The test identifies 92 as the optimal number of topics for the case study corpus.

with some exceptions given by "ridges". These identify certain topic models that appear to be dissimilar. A so-called "ridge" reflects the TWW_k^K statistics for a model with $\hat{K}$ topics. Visually, a topic model with 157 topics has the *relatively* largest TWW_k^K statistics. Even though this models appear to be different from a 92-topic model, the chi-square statistics are small in an absolute sense, i.e., they are not significantly different from zero. As one moves along the $\hat{K}$-topic model axis, one can see that there is no evidence of persistent dissimilarity across models.

Finally, we are able to visualize and evaluate aggregate K topic stability. In other words, we formally test if the corpus displays aggregate K topic stability. Figure 4 shows the $\mathbf{TWW^K}$ statistics associated with **Test 3**. To enhance comparability, we normalize the test statistics by scaling them by their respective means. The scaled means rapidly approach 1.0 — a benchmark that is well below the one-standard deviation cutoff of 1.414.[8] One can see that, relative to the optimal 92-topic model, one cannot reject the hypothesis of topic stability at conventional significance levels as the number of topics increase.

8. A chi-square distributed random variable has a mean equal to the number of degrees of freedom and a variance equal to 2 times the number of degrees of freedom.

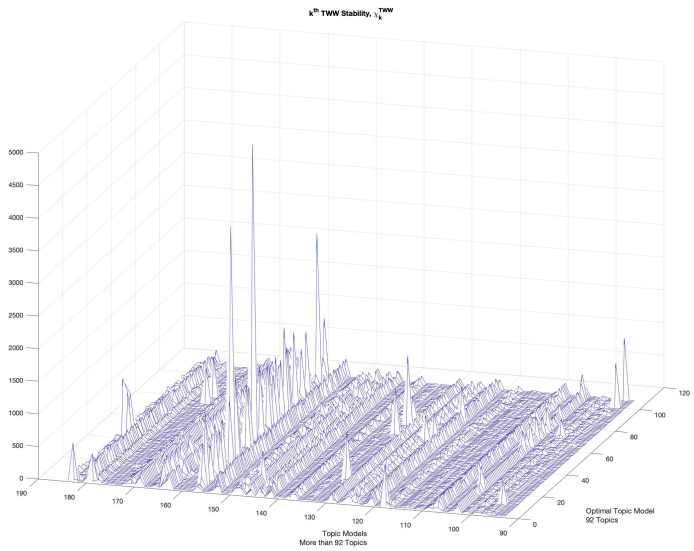

Figure 3: Chi-square Statistics for $k-th$ **TWW Stability** for models ranging from 93 to 200 topics relative to a 92-topic model.

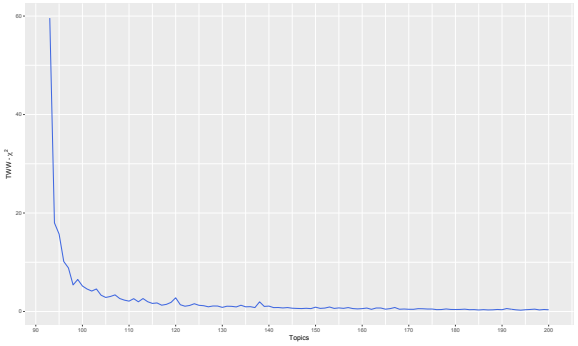

Figure 4: Chi-square statistics for $\mathbf{TWW^K}$ for models ranging from 93 to 200 topics relative to a 92-topic model.

6.3 Addressing Overall Model Adequacy — Tests 4 and 5

Test 4 compares the fitted value from the optimal K-topic model to the informative component derived from $\hat{K}$-topic models. In Figure 5, we show the chi-square statistics $\text{AGG\_STAB}^K_{\hat{I}}$ (Equation (17)) that are obtained for each document in the corpus. The average value is 3.6×10^{-3} and the corresponding median is 7.1×10^{-6}. The $\text{AGG\_STAB}^K_{\hat{I}}$ statistics indicate that the most similar topics from LDA models with $\hat{K}$ topics follow the same distribution as the optimal K-topic model. One can see that a few documents do not

fit particularly well, but they are clearly outliers that do not change our overall assessment of model adequacy.

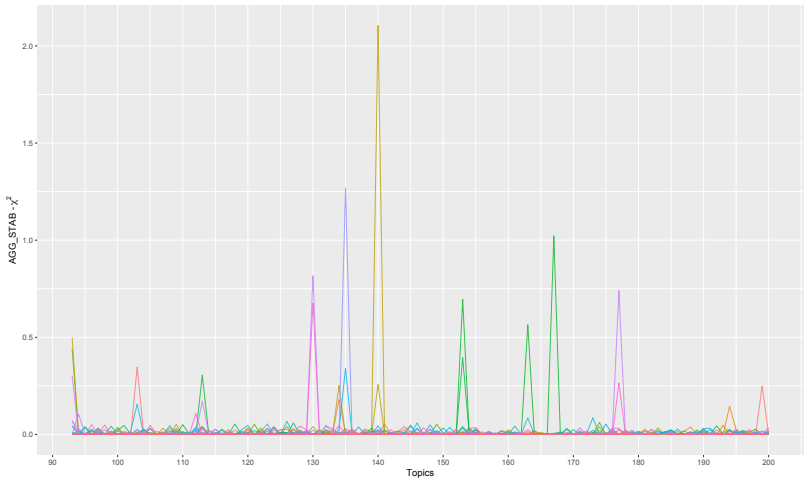

Figure 5: chi-square statistics for $\text{AGG\_STAB}^K_{\hat{f}}$ for above optimality models ranging from 93 to 200 topics. Colors represent different documents in the corpus.

Given that the most similar topics and their corresponding document topic weights are stable for different values of $\hat{K}$, **Test 5** provides a formal testing methodology that compares the relative explanatory power of the informative and uninformative components of different $\hat{K}$-topic models. The F_K^{CORP} statistic given in Equation (18) generally rejects the hypothesis that the uninformative component have significant incremental explanatory power relative to the informative component. Figure 6 depicts the results for the F_K^{CORP} statistic by showing the chi-square statistics for each document across all of the topic models.

7. Simulation Study

We further validate our method by performing a simulation study based on the estimated optimal LDA model with 92 topics. We synthetically construct the true underlying topic structure for our simulations by randomly selecting K_{sim}-columns from **DTW** and K_{sim}-rows from **TWW**. The simulation is then based on a set of synthetic corpora containing a range of topics from 10 to 35 in increments of 5, i.e., $K_{sim} = \{10, 15, 20, 25, 30, 35\}$. For each K_{sim}, we simulate $M = 50$ corpora each of which contains $J = 58$ documents.[9] This generates a total of 300 synthetic corpora on which we re-estimate a battery of LDA models. Specifically, we consider a window of $K_{sim} \pm 20$. This implies, for example, that when the "ground truth" is $K_{sim} = 30$, we estimate a set of LDAs with $k = \{10, \ldots, 50\}$. For $K_{sim} = \{10, 15, 20\}$, we start the estimation at $k = 2$.

For each simulation run, we directly compare our Goodness-of-Fit (GoF) metric *OpTop* with the *perplexity* index (Blei et al., 2003). In Figure 7, we show the results of the simu-

9. The simulation step exploits the classical LDA structure. The full numerical implementation can be found in the function `sim_LDA_data()` in the `R` package `LDATS` (Simonis et al., 2020).

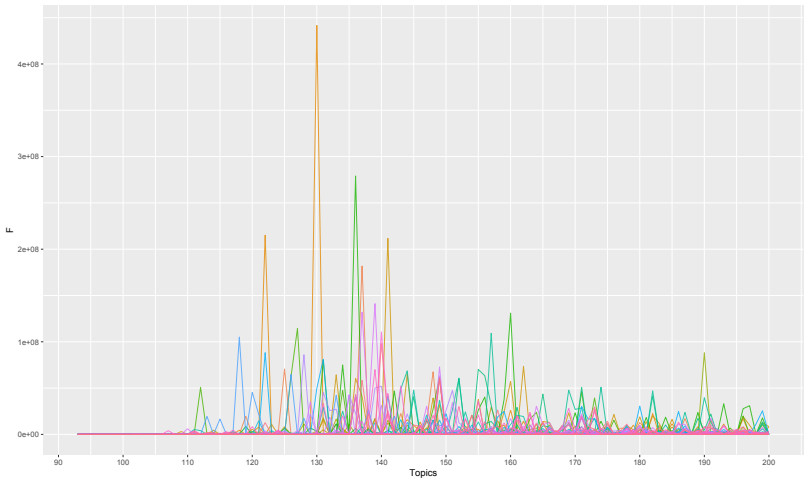

Figure 6: F statistics F_K^{CORP} for above optimality models ranging from 93 to 200 topics. Colors represent different documents in the corpus.

lation study. In each panel, we report the two metrics given as the grand mean computed over the M simulation runs for each LDA specification. The solid black line represents our proposed metric *OpTop*, while the long-dashed black line is the *perplexity* score. The vertical gray dot-dashed line marks the true optimal topic instead (or *ground truth*). The optimal model specification is identified as the point where the reported metrics are minimized. In principle, we would look for a global minimum or a sudden change in the first derivative of the function. As Figure 7 shows, *OpTop* is the only one that "bends" around the true optimal topic. This is particularly clear in Panel (e) where the true optimal topic is $K_{sim} = 30$. Conversely, the perplexity index is monotonically decreasing throughout the range depicted in Figure 7. Note also that the perplexity with respect to k does not bend even after K_{sim}. This indicates that an eventual flattening in the perplexity would occur at a location where the true number of topics would be strongly overestimated. Besides the clear implications regarding the overall statistical performance of the model, this would also imply that a higher dimensional model must be implemented with a demand for much more computational resources.

We acknowledge that *OpTop* does not always identify the exact optimal number of topics. For these cases, the approach still works well as the optimum is typically adjacent or in close proximity to K_{sim}. Table 2 reports the proportion of time *OpTop* is more accurate than *perplexity*. In the majority of cases, *OpTop* outperforms the *perplexity*.

We compute two additional GoF metrics: a density-based method proposed by Cao et al. (2009) and a matrix factorization-based method proposed by Arun et al. (2010) that exploits the KL-Divergence to identify the natural number of topics. In untabulated results, *OpTop* is still more accurate than these two metrics. Overall, the simulation analysis establishes that *OpTop* provides a superior GoF metric relative to other more *ad-hoc* approaches.

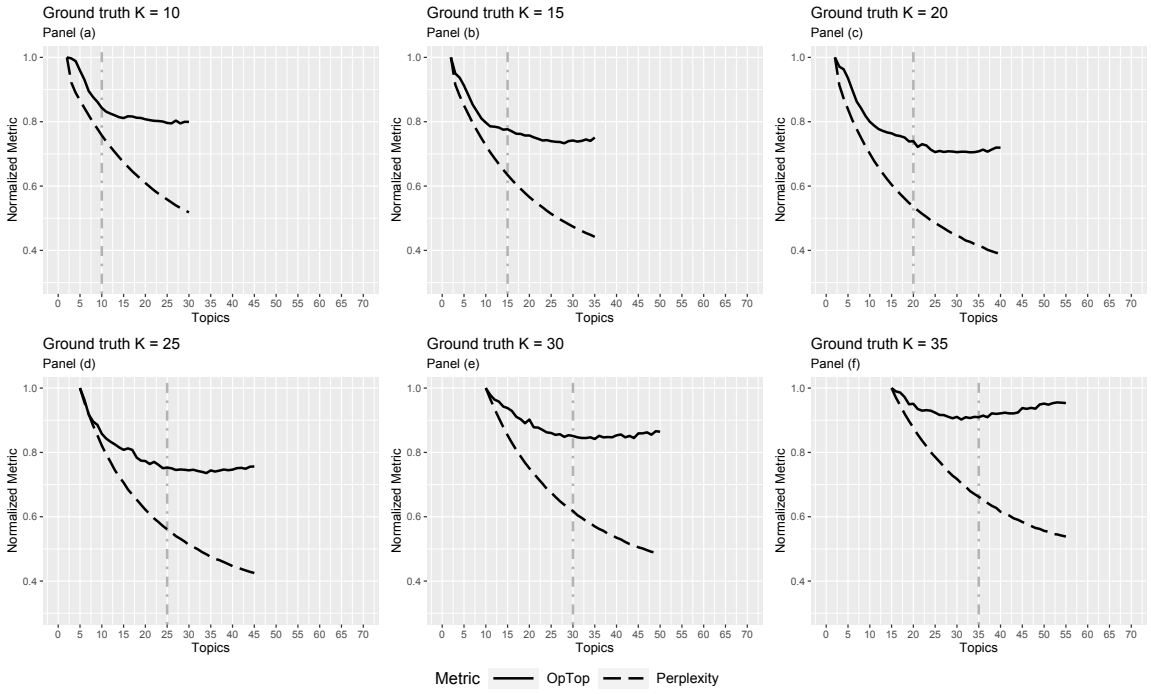

Figure 7: Normalized GoF metrics for a set of LDA models estimated around the true optimal topic K_{sim}. Panels (a), (b), (c), (d), (e), and (f) report the results for $K_{sim} = \{10, 15, 20, 25, 30, 35\}$, respectively. The solid black line represents our metric *OpTop*, the long-dashed line is the *perplexity* (Blei et al., 2003). The vertical gray dot-dashed line marks the true optimal topic.

K_{sim}	*OpTop* vs. *Perplexity*
(1)	(2)
$K_{sim} = 10$	0.84
$K_{sim} = 15$	0.90
$K_{sim} = 20$	1.00
$K_{sim} = 25$	1.00
$K_{sim} = 30$	1.00
$K_{sim} = 35$	1.00

Table 2: Proportion of times when *OpTop* is closer than *perplexity* to the true optimal topic. Column (1) gives the true optimal topic and Column (2) compares *OpTop* with the *perplexity* (Blei et al., 2003).

8. Numerical Implementation

Text processing and management have been carried out with the `R` package `quanteda` (Benoit et al., 2018). LDA models are estimated with the `R` package `topicmodels` (Hornik

and Grün, 2011) which exploits the original `C` code for the VEM fitting implemented by Blei et al. (2003).[10] We estimate 199 consecutive LDA models ranging from 2 to 200 topics over the U.S. presidential inaugural address texts (Peters, 2018) corpus included in the package `quanteda` (Benoit et al., 2018). From this set, we use the procedures introduced in this paper to find the optimal topic specification. The test statistics are calculated using the original code developed in MATLAB. The simulation study relies on the `R` package `LDATS` (Simonis et al., 2020). The authors are currently developing the corresponding `R` package `OpTop` that will calculate all the tests introduced in this work. The package directly interacts with `topicmodels` and the related `LDA_VEM` class (Hornik and Grün, 2011) which provides the estimates for the LDA models.[11]

9. Conclusion

This paper develops a rigorous yet intuitive parametric approach to address the problem of optimal topic identification. Using the fact that Latent Dirichlet Allocation assumes that the vocabulary associated with a corpus can be described by a set of multinomial distributions, we design a chi-square test to identify the optimal number of topics (*Property 1*). We then provide a series of additional chi-square tests to determine i) whether the corpus displays *topic stability* (*Property 2*) and ii) the relative ability of the K most-similar topics to explain actual word choices relative to the $\hat{K} - K$ least similar topics (*Property 3*).

We illustrate the identification strategy using the U.S. Presidential Inaugural Address Corpus as a case study. We determine that a 92-topic model is the optimal specification. As additional topics are added, the optimal specification displays topic stability and the most similar topics explain the actual word choices significantly better than the least similar ones. Moreover, we find that the uninformative topics do not have significant incremental explanatory power. To further validate our findings, we perform a simulation study and we conclude that our proposed statistical approach outperforms other heuristic procedures like the perplexity index.

Acknowledgments

The authors thank Gerard Hoberg for the suggestions and helpful comments on the first draft of this article. The editor, associate editor and referees are also acknowledged for their useful suggestions. The author Francesco Grossetti is also affiliated with the Bocconi Institute for Data Science and Analytics (BIDSA).

10. We use the open source `R` (R Core Team, 2021) programming language for data processing and visualizations. In particular, the former have been carried out with the `data.table` package (Dowle and Srinivasan, 2017) while the latter with `ggplot2` (Wickham, 2009).
11. The package can be found on Github at `https://github.com/contefranz/OpTop`. The development version is available for installation and testing.

References

Alan Agresti. *An introduction to categorical data analysis*, volume 135. Wiley New York, 1996.

Edoardo Maria Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 2008.

Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer, 2010.

Brian Ball, Brian Karrer, and Mark EJ Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, Patrick O Perry, Jouni Kuha, Benjamin Lauderdale, et al. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774, 2018.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

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

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781, 2009.

John E Core, Wayne Guay, and David F Larcker. The power of the pen and executive compensation. *Journal of Financial Economics*, 88(1):1–25, 2008.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

James M Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78(383):628–637, 1983.

M. Dowle and A. Srinivasan. *`data.table`: Extension of `data.frame`. R package version 1.10.4.* `https://CRAN.R-project.org/package=data.table`, 2017.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

Martin Gerlach, Tiago P Peixoto, and Eduardo G Altmann. A network approach to topic models. *Science advances*, 4(7):eaaq1360, 2018.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Robert Gunning. The fog index after twenty years. *Journal of Business Communication*, 6 (2):3–13, 1969.

Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 22, pages 50–57. SIGIR, 1999.

Thomas Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, volume 51, pages 211–218. ACM, 2017.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Kurt Hornik and Bettina Grün. topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Michael Irwin Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, National Technical Information Service, 1975.

David F Larcker and Anastasia A Zakolyukina. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540, 2012.

Feng Li. Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*, 2006.

Feng Li. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102, 2010.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

Tim Loughran, Bill McDonald, and Hayong Yun. A wolf in sheep's clothing: The use of ethics-related terms in 10-k reports. *Journal of Business Ethics*, 89(1):39–49, 2009.

Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. Modeling healthcare data using multiple-channel latent dirichlet allocation. *Journal of Biomedical Informatics*, 60:210–223, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Gerhard Peters. The American Presidency Project, 2018. URL `https://www.presidency.ucsb.edu`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria - `https://www.R-project.org/`, 2021.

Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

Timothy Rubin and Mark Syeyvers. Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*, 2006.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

Juniper L Simonis, Erica M Christensen, David J Harris, Renata M Diaz, and Hao Ye. *LDATS: Latent Dirichlet Allocation Coupled with Time Series Analyses. R package version 0.2.7.*, 2020.

Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

Olivier Toubia, Garud Iyengar, Renee Bunnell, and Alain Lemaire. A topic model for movie choices and ratings. *unpublished University of California Irvine working paper*, 2019.

Hadley Wickham. *`ggplot2`: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics*, volume 16, page S8. BioMed Central, 2015.