Ranking coherence in Topic Models using Statistically Validated Networks

Journal Title
XX(X):2–??
©The Author(s) 2020
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

1 2

Prepared using sagej.cls

Abstract

Probabilistic topic models have become one of the most widespread machine learning techniques in textual analysis. Topic discovering is an unsupervised process that does not guarantee the interpretability of its output. Hence, the automatic evaluation of topic coherence has attracted the interest of many researchers over the last decade, and it is an open research area. The present article offers a new quality evaluation method based on Statistically Validated Networks (SVNs). The proposed probabilistic approach consists of representing each topic as a weighted network of its most probable words. The presence of a link between each pair of words is assessed by statistically validating their co-occurrence in sentences against the null hypothesis of random co-occurrence. The proposed method allows one to distinguish between high-quality and low-quality topics, by making use of a battery of statistical tests. The statistically significant pairwise associations of words represented by the links in the SVN might reasonably be expected to be strictly related to the semantic coherence and interpretability of a topic. Therefore, the more connected the network, the more coherent the topic in question. We demonstrate the effectiveness of the method through an analysis of a real text corpus, which shows that the proposed measure is more correlated with human judgement than the state-of-the-art coherence measures.

Keywords

Text Mining, Probabilistic Topic models, Topic coherence, Statistically Validated Networks

1
2

1 Introduction

The scientific interest in automatic textual analysis has grown dramatically over the last decade. The task of extracting meaningful information from texts has become more important due to the increase in available digital textual data. Indeed, researchers from several disciplines have become increasingly interested in incorporating textual data in their works. Text Mining or Knowledge Discovery from Text (KDT) was first introduced by Feldman and Dagan[1] and refers to the process of extracting high-quality information from text. One of the most critical goals of text mining is the clustering task[2], studied in different research domains such as data mining[3], machine learning[4], and information retrieval[5]. Topic modeling[6] is one of the most popular probabilistic clustering algorithms, since it aims to process extensive collections of texts that are useful for tasks such as classification, novelty detection, summarisation, similarity and relevance judgments.

These models learn topics automatically, from unlabeled documents in an unsupervised way. These topics are called **hidden thematic structure** or latent topics and are typically represented as sets of essential words. Documents are considered as a mixture of topics, where each topic is represented by a probability distribution of words[7]. Thus, these models build latent topics as multinomial distributions of words and the models assume that each document can be described as a mixture of these topics.[8]. Each topic's essential words frequently tend to appear together and (hopefully) are related to the same common theme. Once the models are trained, they provide a framework for humans to understand document collections both directly by "reading" models or indirectly by using topics as input variables for further analysis[9]. The Latent Dirichlet Allocation

Corresponding author:

(LDA) is one of the most popular topic models and the state-of-the-art unsupervised machine learning technique for extracting thematic information (topics) from a collection of documents.

Indeed as highlighted by Boyd-Graber et al.[9], LDA plays an essential role in the analysis of historical documents, scientific documents, fiction, poetry and literature. The main obstacle in topic detection models is that not all the estimated topics are of equal importance and not all correspond to genuine domain themes. Some of the topics can be a collection of irrelevant words or unchained words representing insignificant themes.

Often, in qualitative studies, the goal is to find meaningful and interpretable topics. Researchers usually use top-N words with the highest probability given a topic[10–13], and employ humans to obtain an interpretability score. Indeed, topic discovering algorithms do not automatically provide a way to interpret their output. For instance, Chang et al.[8] state that "Although there appears to be a longstanding assumption that the latent space discovered by topic models is meaningful and useful, evaluating such assumptions is difficult because discovering topics is an unsupervised process". Moreover, Hoyle et al.[14] highlight that automated evaluation metrics often suffer from inconsistency. Therefore, it would be desirable to fully automize the process by introducing a metric that automatically ranks learned topics closely matching human judgments. This challenge motivated recent research on topic quality metrics that closely match human judgement. Within this framework, quantifying the coherence of a set of words plays a central role[10–13,15–17] .

In topic models, a topic can be viewed as a set of words that frequently co-occur in the same documents, which is very similar to latent word groups (or communities)[18] in the word network. Since words that frequently co-occur in the same sentences are closely connected in the semantic space, they tend to appear in the same document.

This paper proposes a new topic coherence measure based on the

construction and analysis of Statistically Validated Networks (SVNs) of words[19]. Specifically, the method builds a co-occurrence network for each topic whose most probable words are the nodes. We set a link between two nodes (words) in each network if their co-occurrence in sentences is statistically significant. We claim that these links carry relevant information about the structure of the topic, i.e., the more connected the network, the more semantically coherent the corresponding topic. Therefore, we propose to use connectivity measures on the SVN of words to build a metric of topic coherence.

The main contributions of this paper are: i) to define a new coherence measure (Coh_{SVN}) based on a rigorous statistical model that approximates human ratings better than state-of-the-art methods; ii) to filter out marginal associations of words and to facilitate the graphical representation and interpretation of the obtained topics through Statistically Validated Networks (SVNs)[19].

1.1 Organization of the paper

The paper is organized as follows: Section 2 describes the background and reviews related works. In Section 3, we describe the proposed coherence model, while we report a real-world application of the method in Section 4. Finally, in Section 5, we draw our conclusions and propose ideas for future development.

2 Background and related works

The main idea of topic modeling is to create a probabilistic generative model for a corpus of text documents. A probabilistic topic model is a type of generative model that aims to learn the latent semantic structure of a corpus. Probabilistic topic models reduce the complex process of document generation to a small number of probabilistic steps by

assuming exchangeability, because only word occurrence information (i.e., frequencies) is considered.

The first probabilistic topic model was the Probabilistic Latent Semantic Analysis (pLSA), introduced by Hofmann[20]; unfortunately, the model does not provide any probabilistic model at the document level. Then, Blei[6] proposed the The Latent Dirichlet Allocation (LDA) model as an extension of the pLSA, introducing a Dirichlet prior on mixture weights of topics per document. The name of the model incorporates its main features. Specifically, the term *Latent* indicates that the model involves probabilistic inferences for extrapolating missing probabilistic pieces of the generative story from texts. The term *Dirichlet* recalls that the model uses Dirichlet parameters to encode sparsity. Finally, the name includes the word *Allocation* since the Dirichlet distribution encodes the prior probability for each document's allocation of the topics[9]. In these models, documents are described as random mixtures over latent topics, where a distribution of words characterizes each topic[6]. The words of the documents are the observed variables, whereas the topic structures are the hidden variables. The problem of inferring the hidden topic structure from the documents consists in computing the posterior distribution of topic structures, that is, the conditional distribution of the hidden variables given the documents[7].

Recently, many other probabilistic topic models that consider topic correlations were proposed, such as the correlated topic model (CTM– see Blei and Lafferty[21]), the Pachinko allocation model (see Li and McCallum[22]). Other works extend probabilistic topic models focusing on the evolution of topics over time, such as the dynamic topic model (DTM)[23], or introducing word embedding representation–the embedded topic model (ETM) by Dieng et al.[24].

Finally, neural topic models represent a broader set of related models.

These mainly focus on improving topic modeling inference through deep neural networks (see Srivastava and Sutton[25]).
Finally, Blei (2012)[7] and Boyd-Graber et al. (2017)[9] provide comprehensive reviews of probabilistic topic models.
Among these models, we applied our coherence measure to the LDA model, since it represents a benchmark in the topic modelling community, for comparison with its various extensions. However, it is worth highlighting that the proposed measure applies to any topic model.

2.1 Literature review

Evaluating the quality of the latent spaces provided by topic models is a difficult challenge because discovering topics is an unsupervised process that gives no guarantees on the interpretability of its output. In text mining, the problem of semantic evaluation has attracted much interest breaking down the research into coherence measures[17]. There is no gold-standard list of topics to compare against for every corpus. Thus, a technique for evaluating the outputs of topic models could be employed on gathering exogenous data. In this section, we discuss previous work on the topic evaluation.
For many years, the primary way to evaluate the quality of a topic model was to measure the log-likelihood of a held-out test set[6,26]. The held-out likelihood consists in density estimation on a collection of unseen documents given a training set. The most commonly used measure based on the held-out method is the perplexity, a monotonically decreasing function of likelihood:

$$perplexity(D) = exp\Big\{ - \frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d} \Big\},$$

where D is the collection of documents, N_d is the number of words in document d, and $p(\mathbf{w}_d)$ is the marginal distribution of document d,

following the notation used in previous section. A lower perplexity score indicates better generalization performance.

However, Chang et al.[8] showed that the perplexity on held-out test set emphasizes *complexity* rather than *interpretability*, which is the property users are mostly interested in. In their work, they fit three different topic models to two corpora and demonstrated that the perplexity scores are negatively correlated with human ratings. In other words, such measure is useful for evaluating the predictive performance of the model, but it do not address the more explanatory goals of topic modeling. Indeed, topic models are mainly used to organize, summarize and help users to explore large corpora, while evaluating the predictive performance of the model is a completely different task. Therefore, there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation. In recent years, many methods have been proposed for assessing topic coherence. The approaches can be split into two categories: qualitative methods and quantitative methods. Qualitative methods are less common than quantitative, since require the use of human resources for topic assessment, and are time-consuming. Quantitative approaches, on the other hand, seek to automate the whole evaluation process trying to replicate human judgment.

2.2 Qualitative methods

Chang et al.[8] proposed the task of *word intrusion* to create a formal setting where humans can evaluate the latent space of a topic model. This task allows for an evaluation of whether a topic has human-identifiable semantic coherence or not. In the *word intrusion* task, the subject is presented with six randomly ordered words, and the task of the user is to find the word which is out of place or which does not belong with the others, i.e., the *intruder*.

In 2018, Morstatter and Liu[27] proposed a modified version of the word

intrusion task, named *Model Precision Choose Two*. As in the word intrusion task, they propose to form a list with the top (most likely) five words from a topic and to inject one low-probability word from the same topic into the list. The critical difference with word intrusion is that they ask the annotators to select *two* intruded words from the six. The intuition behind this experiment is that the annotators' first choice will be the intruded word, just as in[8]. However, their second choice is what makes the topic's quality clear. In a coherent topic, the annotator will not be able to distinguish a second word as all of the words will appear similarly coherent.

2.3 Quantitative methods

The qualitative methods are time consuming since they require the manual annotations of humans. In the last decade, researchers have proposed to fully automating the process by introducing a metric that allows to automatically rank learned topics. One of the first automated measure was proposed by AlSumait et al. in 2009[15]. They introduced an approach to **automatically** rank the LDA topics based on their semantic importance and, eventually, to identify junk and insignificant topics. Their idea is to measure the amount of "*insignificance*" that an inferred topic carries in its distribution by measuring how "different" the topic distribution is from a "*junk*" distribution. In the same work, AlSumait et al. proposed three definitions of Junk and Insignificant (J/I) topic distribution, namely: i) the Uniform Distribution Over Words *(W-Uniform)*, ii) the Vacuous Semantic Distribution *(W-Vacuous)* and iii) the Background Distribution *(D-BGround)*. Finally, to quantify the difference between an estimated topic and a J/I distribution, three different distance measures are employed, namely: Kullback-Leibler (KL) Divergence; Cosine Dissimilarity; and Correlation Coefficient.

Later, Wang et al.[28] proposed a re-ranking algorithm to select "significant" topics by topic similarity calculation. Specifically, each topic is represented as a probability distribution $p(w_i|z_j)$ over words. To compute the distance between word-topic distributions they employed the Jensen-Shannon distance (a symmetrised extension of the KL divergence) :

$$Dist(z_i, z_j) = \frac{1}{2}[KL(z_i||z_j) + KL(z_j||z_i)].$$

Finally, for each topic i, they computed the average distance between i and all the other topics, and they sort the average distance for each topic in a queue. The last element in the queue is ranked the highest.

In the framework of topic quality evaluation, many relevant works make use of the top-N most probable words (rather than using the entire word-topic distribution), and they assess pairwise semantic cohesion among them through their co-occurrences provided by the dataset or external sources.

The general idea is to compute the mean of the sum of the pairwise scores of the top-N words that most contribute to describing the topic:

$$Coherence = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} score(w_i, w_j).$$

One of the best-known topic quality measures based on the top-N words was proposed by Newman et al.[29]. They introduced for the first time, a model that uses external text data sources, such as Wikipedia and Google hits, to predict human judgements.

Specifically, Newman et al.[29] measured co-occurrence of word pairs, taken from the list of the ten most probable words in a given topic, using two huge external text datasets: all articles from English Wikipedia and the Google n-grams data set. Specifically, they identify a co-occurrence of words w_i and w_j if they occurred together in a 10-word window of

any Wikipedia article. Similarly, they identify a co-occurrence of the two words according to Google n-grams if they both appear in any of the existing 5-grams. Finally, they measure the score of association between word pairs through the Pointwise Mutual Information (PMI)[30]:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}, \tag{1}$$

where $p(\cdot)$ is the relative frequency of a word and $p(\cdot, \cdot)$ is the relative frequency of the co-occurrence of two words, while ϵ is a smoothing term. This measure is also called *UCI*.

Minmo et al.[31] pointed out that "bad" topics can be categorized into three definitions:

- *Chained*: every word is connected to every other word through some pairwise word chain, but not all word pairs make sense.
- *Intruded*: either two or more unrelated sets of related words, joined arbitrarily, or an otherwise good topic with a few "intruder" words.
- *Random*: no clear, reasonable connections between more than a few pairs of words.

In their work, the authors suggest that these poor-quality topics could be detected using metrics based on word co-occurrences within the documents.

They proposed to use an asymmetrical confirmation measure, $UMass$, between top word pairs (smoothed conditional probability), where the estimations of word probabilities are based on their frequencies in the original documents used to train the algorithm on the topics:

$$UMass(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_j)}, \tag{2}$$

where $D(w_i)$ is the *document frequency of word*, (i.e., the number of documents that contains w_i, and $D(w_i, w_j)$ is *co-document frequency* (i.e.,

the number of documents containing both words). Note that Eq. 2 is equal to the empirical conditional log-probability $\log p(w_i|w_j) = \log \frac{p(w_i,w_j)}{p(w_j)}$ smoothed by adding one to $D(w_i, w_j)$, where $p(w_i) = \frac{D(w_i)}{M}$. Therefore, the score function is not symmetric as it is an increasing function of the empirical probability $p(w_j|w_i)$, where the probability of w_i is higher than the word w_j, given a topic. Therefore, this score measures how much (within the words used to describe a topic) a common word, w_i, is on average a good predictor for a less common word, w_j.

Another important contribution was given by Lau et al.[10] who proposed to use the Normalized Pointwise Mutual Information (NPMI)[30] of word pairs in the automated methods of word intrusion and observed coherence:

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log\left[p(w_i, w_j) + \epsilon\right]}, \tag{3}$$

where $p(\cdot)$ and $p(\cdot, \cdot)$ are defined as for PMI. The NPMI ranges between (-1,+1) resulting in -1 (in the limit) for never occurring together, 0 when they are distributed as expected under independence, and +1 (in the limit) for complete co-occurrence.

Aletras and Stevenson[12] proposed a method for determining topic coherence using the distributional similarity between the n most likely words of the topic. Representing each word as a vector, let $\overrightarrow{w_1}$,$\overrightarrow{w_2}$,..., $\overrightarrow{w}_n$ denote the vectors of the top n most probable words in a topic. The authors also assume that each vector consists of N elements (the size of the Vocabulary) and $\overrightarrow{w}_{ij}$ is the jth element of vector $\overrightarrow{w_i}$. The semantic space was created using Wikipedia as a reference corpus and a window of ± 5 words. Then they compute the similarity between words using three measures:

- Cosine similarity:

$$Cos(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{\overrightarrow{w_i} \cdot \overrightarrow{w_j}}{\|\overrightarrow{w_i}\| \|\overrightarrow{w_j}\|}$$

- Dice coefficient:

$$Dice(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{2\sum_{k=1}^{N} \min(\overrightarrow{w_{ik}}, \overrightarrow{w_{jk}})}{\sum_{k=1}^{N} (\overrightarrow{w_{ik}} + \overrightarrow{w_{jk}})}$$

- Jaccard coefficient:

$$Jaccard(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{\sum_{k=1}^{N} \min(\overrightarrow{w_{ik}}, \overrightarrow{w_{jk}})}{\sum_{k=1}^{N} \max(\overrightarrow{w_{ik}} + \overrightarrow{w_{jk}})}.$$

Then, the coherence of topics is constructed by the mean of all pairwise scores. Each of these measures estimates the distance between a pair of words in a topic and produce a topic cohesion measure based on distributional semantics. Roder et al.[17] proposed a framework that allows for the construction of existing word-based coherence measures as well as new ones, by combining elementary components. They conducted a systematic search of the space of coherence measures for the evaluation and they identified a complex combinations (named CV) as the best performers on their test corpora.

Omar et al.[32] quantitatively describe topics via normalized mean values of pair-wise word similarities. They used two types of word similarities, namely, thesaurus and local corpus-based as the descriptive features of a topic, and performed topic classification by using the represented topics as input and a binary 0-1 human ratings.

Some of the latest work in the field was produced by Nikolenko et al.[16]: they highlighted that the topic coherence defined by Minmo et al.[31] is able to consistently identify bad topics (i.e., topics with poor coherence) but does not perform well in identifying good ones (i.e., topics with a high degree of coherence). To cope with this problem, Nikolenko et al.[16] proposed tf-idf (term frequency - inverse document frequency) coherence

as a modification of Mimno's coherence metric that accounts for the informative content of the topics.

Their idea is to introduce $tf\text{-}idf$ scores instead of the number of co-occurrences in order to construct their measure. The $tf\text{-}idf$ value, as defined by Salton and Buckley[33], increases proportionally to the number of times a word appears in a document and is inversely proportional to the number of documents in the corpus that contain that word. This measure privileges the words that not only frequently occur in a given text, but that also occur rarely in other texts. Thus, a coherence metric with $tf\text{-}idf$ scores penalizes co-occurrence of common words that have low discriminative power. The measure for a given topic is defined as follow:

$$C_{tf\text{-}idf}(w_i, w_j) = log \frac{\sum\limits_{d:w_i,w_j \in d} tf\text{-}idf(w_i, d) tf\text{-}idf(w_j, d) + \epsilon}{\sum\limits_{d:w_i \in d} tf\text{-}idf(w_i, d)},$$

where ϵ is a smoothing count usually set to either 1 or 0.01, while the $tf\text{-}idf$ metric is computed with augmented frequency:

$$tf\text{-}idf = tf(w, d) \cdot idf(w, d),$$

where

$$tf(w, d) = \Big(\frac{1}{2} + \frac{f(w, d)}{\max_{w^* \in d} f(w^*, d)} \Big),$$

$$idf(w, d) = log \frac{|D|}{|\{d^* \in D : w \in d^*\}|}.$$

3 Methods

In this section, we propose a new coherence measure to evaluate the interpretability of the top words of a topic. Our method consists in building a co-occurrence network for each topic whose most probable words (according to the estimated topic model) are the nodes. The weights of links are calculated as the number of sentences in which the connected words co-occur. In each network, we identify the links whose weight is statistically significant, i.e., those that cannot be explained in terms of random co-occurrences of words in the sentences. Although several measures in the literature have already considered co-occurrence between words as a measure of association, none has undertaken a statistical approach based on hypotheses testing to assess whether the co-occurrence obtained between two words can be attributed to chance or whether these links carry relevant information about the structure of topics. To do this, we exploit Statistically Validated Networks.

3.1 Statistically Validated Networks

In recent years, many complex systems have been represented by bipartite networks[34–36]. The Statistically Validated Network, introduced by Tumminello et al.[19], is an unsupervised method to statistically test the significance of each link of a projected weighted network as obtained from a multipartite network. It is an unsupervised method that introduces a system of hypotheses for link testing when a multipartite network is projected into a set of nodes. The idea is to represent text data as a bipartite network, Fig1, in which the set of nodes S is made by the sentences of corpus and the other set of nodes W is made by a list of words associated with a given topic. A link is set between a word and a sentence if the word belongs to that sentence. Therefore, projecting the set of words, the resulting network is a word-co-occurrence network[18,37].

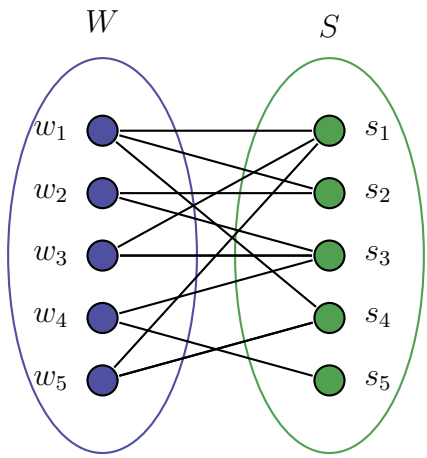


Figure 1. Bipartite network where S is the set of corpus sentences and W is the set of topic words.

To take into account the heterogeneity of the set of sentences, a suitable system of hypotheses is introduced. The hypothesis test is constructed as follows. Let us consider a corpus made of N sentences, then consider two words, say, w_i and w_j, and indicate with X_{ij} the times they appear in the same sentences. We are interested in validating the co-occurrences of the words w_i and w_j statistically against a null hypothesis of random co-occurrence that accounts for the heterogeneity of the considered words, that is, the total number of times they appear individually in the text, N_i and N_j, respectively. The probability distribution that describes the random co-occurrence is the hypergeometric distribution, according to which, the probability of observing X_{ij} co-occurrences is given by

$$\text{pmf}_H(X_{ij}|N, N_i, N_j) = \frac{\binom{N_i}{X_{ij}}\binom{N-N_i}{N_j-X_{ij}}}{\binom{N}{N_j}}$$

where parameters N_i and N_j naturally allow for the incorporation of the aforementioned heterogeneity of words in the null hypothesis.

The Hypergeometric distribution describes the probability mass function

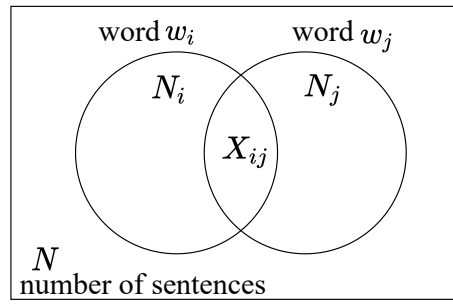


Figure 2. Venn Diagram showing the overlap of two words

under the null hypothesis in which the probability of co-occurrence between words is conditioned by their marginals, i.e., their individual occurrences.

The distribution introduced can be used to test the presence of an excess of co-occurrence between any pair of words, w_i and w_j. Indeed, assuming that the actual co-occurrences of these words is N_{ij}, then the probability that a value larger than or equal to N_{ij} is observed by chance, according to the null hypothesis, is:

$$p_v(N_{ij}|N_i, N_j, N) = \sum_{X=N_{ij}}^{min(N_i,N_j)} \frac{\binom{N_i}{X}\binom{N-N_i}{N_j-X}}{\binom{N}{N_j}}. \quad (4)$$

To claim that the number of co-occurrences, N_{ij}, between words is too large to be consistent with the null hypothesis of random co-occurrences, we shall set a threshold α of statistical significance. However, since we are facing multiple and dependent comparisons, errors of the first kind are a real issue. Therefore, we use the conservative Bonferroni correction[38] for multiple hypothesis testing. The correction states that given a univariate threshold of statistical significance, α, then the threshold corrected for multiple hypothesis testing is $\alpha_T = \frac{\alpha}{T}$, where T is the total number of performed tests, be they dependent or otherwise. The advantage of the Bonferroni correction is that it provides a very strict control of the Family

Wise Error Rate even when tests are dependent, as they are in this case, since the same word appears in many tests.

3.2 Coherence based on SVNs

In this section, we describe how to construct the new coherence measure, Coh_{SVN}, which makes use of Statistically Validated Networks as combined with different word similarity indices. Specifically, our algorithm can be summarised in the following 5 steps, also sketched in the diagram reported in Fig. 3:

(A) Estimate a topic model, and extract the top-m words from each estimated topic;

(B) Represent each topic as a Statistically Validated Network of words;

(C) Evaluate each link's importance, $Imp(w_i, w_j|z_k)$ by considering the strength of the association between word pairs and the relative relevance of each word in the topic;

(D) Compute a global measure of coherence, Coh_{SVN}, for each topic network;

(E) Produce the final ranked list of topics, by sorting them in decreasing order of coherence.

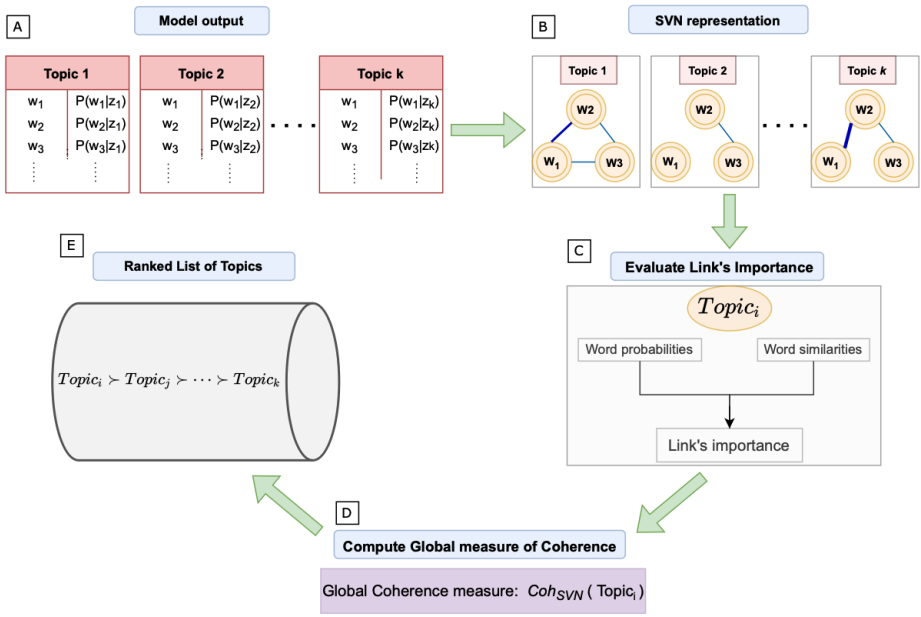


Figure 3. Diagram describing the 5 steps of the algorithm.

Regarding the first step, the specific topic model used, the parameter tuning and the choice of the optimal number of topics lay outside the scope of this paper. Relevant insights on these subjects can be found in references[39–42]. The estimation of the LDA model provides a list of K latent topics, each one described by an ordered list of words. So, to conclude the first step, we select the m most probable words*.

To build the SVN of a given topic, $\frac{m(m-1)}{2}$ statistical tests (against the null hypothesis of random co-occurrence) are performed, one for each pair of words.

The results are K weighted Statistically Validated Networks with m nodes and a number of links equal to the number tests that rejects the null hypothesis of random co-occurrence at a given level, α, of statistical

*In the present application, we follow the standard approach of setting $m = 10$.

significance, after the Bonferroni correction for multiple hypothesis testing. An example is shown in Figure 4.

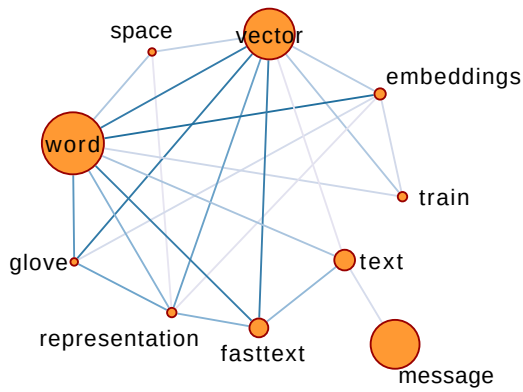


Figure 4. Statistically Validated Network of an artificial topic.

The size of each node i in Figure 4 is proportional to the probability $P(w_i|z_k)$ that the corresponding word w_i appears in the topic z_k, while the opacity of each link is proportional to the strength of the association between the linked words.

To compute the strength of each validated link, we use corpus-based word similarities within distributional contexts. Specifically, let N denote the total number of sentences in the corpus, N_i and N_j the occurrences of words w_i and w_j, respectively, in the sentences of the corpus, and N_{ij} their co-occurrence. To calculate word similarities we use four metrics already used in other studies. Specifically:

- S_1: Jaccard similarity index[43]

$$J(w_i, w_j) = \frac{N_{ij}}{N_i + N_j - N_{ij}} \tag{5}$$

- S_2: Dice-Sorensen coefficient[44] †

$$Dc(w_i, w_j) = \frac{2N_{ij}}{N_i + N_j} \tag{6}$$

- S_3: Sokal and Sneath coefficient[45]

$$SS(w_i, w_j) = \frac{N_{ij}}{2N_i + 2N_j - 3N_{ij}} \tag{7}$$

- S_4: Fowlkes–Mallows index[46]

$$FM(w_i, w_j) = \sqrt{\frac{N_{i,j}^2}{N_i \cdot N_j}}. \tag{8}$$

Furthermore, we also consider three metrics that are tightly related to the SVN method. These metrics are:

- S_5: Similarity based on the Pearson's correlation coefficient $\rho(w_i, w_j)$:

$$D_\rho(w_i, w_j) = \frac{1}{2}\left[1 + \rho(w_i, w_j)\right] \tag{9}$$

where

$$\rho(w_i, w_j) = \frac{N_{ij} - \frac{N_i N_j}{N}}{\sqrt{N_i(1 - \frac{N_i}{N})N_j(1 - \frac{N_j}{N})}}.$$

Since the expected value of the Hypergeometric distribution $H(X|N, N_i, N_j)$ is $\frac{N_i N_j}{N}$ and the variance $\mathbb{V}[X] = \sigma_H^2 = \frac{N_i N_j}{N}\frac{N-N_i}{N}\frac{N-N_j}{N}$, it turns out that $\rho(w_i, w_j)$ is proportional to the Z-score of N_{ij} under the null hypothesis‡.

†Notice that it is equivalent to F1 score.

‡The constant of proportionality is $N^{-\frac{1}{2}}$.

- S_6: Normalized logarithmic robustness $\tilde{R}$

$$\tilde{R}(w_i, w_j) = \frac{\log_{10}(N) - \log_{10}(N^*|w_i, w_j)}{\log_{10}(N) - \log_{10}(n^*|w_i, w_j)}, \tag{10}$$

where

$$N^* = \min\{N : p_v(N_{ij}) < \frac{\alpha}{T}\},$$

is defined as the minimum number of sentences needed in the corpus to validate the co-occurrence between w_i and w_j. While,

$$n^* = \min\{N : p_v(N^*_{ij}) < \frac{\alpha}{T}\}$$

is the minimum value of sentences needed to validate the co-occurrence between w_i and w_j assuming a perfect co-occurrence, $N^*_{ij} = \min(N_i, N_j)$.

- S_7: Similarity based on the normalized p-value $\tilde{p}_v$

$$\tilde{p}_v(w_i, w_j) = 1 - \frac{p_v(N_{ij}|N_i, N_j, N)}{\alpha/T}, \tag{11}$$

where $p_v(N_{ij}|N_i, N_j, N)$ is computed following Eq.4.

All of the proposed similarity measures, $\{S_1, \ldots, S_7\}$, take values in the range $[0, 1]$ where 0 indicates two totally unrelated words, while 1 indicates two perfectly associated words.
Given a validated link between two words, say w_i and w_j, belonging to the topic z_k, we define the link's importance $Imp(w_i, w_j|z_k)$:

$$Imp(w_i, w_j|z_k) = \sqrt{P(w_i|z_k)P(w_j|z_k)}\; S_h(w_i, w_j), \tag{12}$$

where S_h is one of the similarity function described above: $\{D_\rho, \tilde{R}, \tilde{p}_v, J, Dc, SS, FM\}$. The importance of a validated link (Eq. 12), between w_i and w_j give a topic z_k, takes into account two components:

- the relative relevance of w_i and w_j within z_k:

$$\sqrt{P(w_i|z_k)P(w_j|z_k)};$$

- the strength of the association between w_i and w_j:

$$S_h(w_i, w_j), \;\; h = 1, \ldots, 7.$$

The conditional probabilities $P(w_i|z_k)$ and $P(w_j|z_k)$ reflect the relevance of words w_i and w_j, respectively, within the topic z_k. That is to say, words with a higher probability are more relevant within a topic. Therefore, the more relevant two terms, the more important the validated link between them. We decided to use the geometric mean of $P(w_i|z_k)$ and $P(w_j|z_k)$ as aggregating function to reduce the impact of the distribution's tails. As regards to $S_h(w_i, w_j)$, it measures the association between w_i and w_j. Intuitively, the higher the association between two words, the greater the importance of the link between them.

Note that, if w_i and w_j exhibit a "perfect" co-occurence, i.e., $N_i = N_j = N_{ij}$, then $S_h(w_i, w_j) = 1$ and the link's importance reduces to $Imp(w_i, w_j|z_k) = \sqrt{P(w_i|z_k)P(w_j|z_k)}$, that is, the geometric mean of the words probabilities, given the topic, provided by the model.

Finally, we define the **global coherence** measure of a topic, z_k, as:

$$Coh_{SVN}(z_k) = \frac{\sum\limits_{w_i \neq w_j, \in \mathcal{L}} Imp(w_i, w_j|z_k)}{\sum\limits_{w_i \neq w_j, \in \Omega_k} \sqrt{P(w_i|z_k)P(w_j|z_k)}}, \tag{13}$$

where $\mathcal{L}$ is the set of word pairs linked in the SVN, while Ω_k is the set of all possible $m \cdot (m-1)/2$ word pairs for topic z_k.

In Eq.13, the denominator represents the coherence of a perfectly coherent topic, that is a fully connected network where all the pairwise word similarities are maximized, i.e. $S_h(w_i, w_j) = 1 \; \forall w_i, w_j \in \Omega_k$. Thus, $Coh_{SVN}(z_k)$ ranges in the set $[0, 1]$, where the minimum value indicates a totally incoherent and unintelligible topic, while a value of 1 represents a perfectly coherent topic.

Measure $Coh_{SVN}(z_k)$ allows us to rank topics in decreasing order of coherence, which completes the fifth (and final) step of the procedure presented in this section.

4 Experimental evaluation

4.1 Dataset and pre-processing

We evaluated our estimator of topic quality on a dataset of articles extracted from the *New York Times*, which was already analysed by[47]. The dataset (NYTd from now on) consists of 8,764 articles of the *New York Times*, which appeared between April and July 2016[§].

In particular, we decided to consider a reduced version of this dataset, obtained by removing all the articles with fewer than 20 total words (Hong and Davison[48] discuss how short documents can confuse topic modeling algorithms), and taking a random sample of size 1,000 out of those remaining.

The following step is to perform data preprocessing in order to reduce noise from the data. The preprocessing usually consists of tasks such as tokenization, filtering, and either lemmatization or stemming. Tokenization means transforming sentences in a list of words, called token, and the filtering step implies removing all punctuation and numbers. Lemmatization and stemming are two text normalization techniques for Natural Language Processing. The first one is the process

[§]https://www.kaggle.com/nzalake52/new-york-times-articles

of finding the base or dictionary form of a word, called *lemma*, with the aim to remove only inflectional endings considering morphological analysis as meaning and context. Instead, stemming is a method to convert words into their root form by cutting the suffix or prefix from the word. Comparing the lemmatization and stemming methods, we opted for the lemmatization. Stemmed words, in general, are very complicated to interpret, since roots of words were insufficient to discriminate among alternative meanings[49]. For instance, the word *better* has *good* as its lemma, but this link is missed by stemming. We removed urls, mails, punctuation and numbers from the texts through the `Python regex` function. Then, we transformed uppercase letters into lowercase letters and removed accents. Furthermore, we used the `gensim` library to construct compound words, such as *United_States* or *North_Korea*, and `spaCY`, an open-source natural language processing library for Python, to split up sentences. Finally, we removed i) infrequently used words (i.e. appearing only once per document); and ii) redundant words (a rule of thumb is to remove terms appearing in more than 80% of the documents). As a matter of facts, infrequently used terms will not contribute much information about topics, while discovery and removing them may greatly reduce the size of the vocabulary[50]. Equally, it has been shown that redundant words appearing frequently do not convey any meaningful message for topic modeling[51].

The original corpus dictionary, as directly obtained from the 1,000 articles, consisted of $28,104$ tokens, whereas the final corpus (after data preprocessing) included $8,770$ tokens.

The LDA model was trained in `R` setting 50 topics[52], then we randomly extracted 30 of them for human judgment evaluation.

We have chosen to use only part of the group of estimated topics due to time constraints. Indeed, we structured the questionnaire so that each annotator took, on average, 15 minutes to complete their task, assuming

an average response time of about 30 seconds per topic. This issue is crucial for maximising the quality of the answers obtained; in fact, a questionnaire which takes too long to be completed entails the risk of receiving unreliable answers as the respondent's focus drops.
Finally, we prepared graphical representations of the networks of topics using `Cytoscape` software.[¶]

4.2 Coherence-based topic annotations

To obtain high-quality ratings, the survey was structured in two steps. During the first step, which we call "pilot", 23 PhD students from the Department of Economics, Business and Statistics at the University of Palermo, Italy, were brought in. We provided them with 32 topics (consisting of 10 words each) to be evaluated on a 5-point scale where 5="coherent" and 1="not coherent". Among topics, 30 were genuine topics according to the LDA model as applied to the New York Times dataset, and the remaining two were synthetic (control) topics. The first synthetic topic included a group of unrelated words that formed a meaningless and incoherent topic, $z_{31} = \{$Lasagna; Finance; Jeans; Buddhist; Pokemon; Drive; Molecule; Sound; Chess; Revolver$\}$. Instead, the second synthetic topic included perfectly coherent words that formed a strongly coherent topic, $z_{32} = \{$Black; White; Red; Green; Pink; Purple; Brown; Yellow; Grey; Blue$\}$.
We also provided textual guidelines on how to judge whether a topic was coherent or incoherent. In addition to showing several examples of such topics we provided the following preliminary instructions to the respondents.

[¶]https://cytoscape.org

Guidelines

Topic modeling consists of the automatic extraction of groups of words, called *topics*, from a collection of texts. For a topic to be "coherent", it must make sense and be interpretable. This means that the topic's words must:

1. be related to each other
2. belong to the same theme

An automatic procedure for the identification and evaluation of topics is reliable if the topics identified are coherent and interpretable for humans. This is why we are asking you to be part of a benchmark sample of individuals to test the effectiveness of a new topic modeling algorithm we are working on. Therefore, we ask you to rate the coherence of specific topics on a scale of 1 to 5. For example, you can give a topic a low mark if you find few links between the words in it, the mark increases as the number of linked words increases.
It is not always easy to evaluate a list of words, especially if some of them are unfamiliar or belong to a language other than yours (in this case, English). We ask you, *PLEASE*, we ask you to translate any words or nouns you do not know to give as informed a mark as possible.
You will notice that some topics share one or more words; this is not a problem! The topics are not related to each other, so each topic must be evaluated individually. There is no right or wrong answer, since we aim to collect your subjective opinion.

The role of the pilot was to assess the topic annotators' ability in understanding their assigned task. We also investigated which

improvements were necessary in letting annotators deepen their comprehension of the meaning of "coherence".

The most critical issue in the pilot was to investigate whether an odd scale was appropriate. Thus, we studied the relationship between the percentage of neutral answers given by an annotator (i.e. giving a grade of 3) and their probability of failing at least one control topic evaluation.

Table 1. Relationship between giving neutral answers and failing at least one control topic evaluation

	Fail control		
Neutral responses	No	Yes	Total
$\leq 30\%$	14	1	**15**
$> 30\%$	2	6	**8**
Total	**16**	**7**	**23**

Table (1) shows that these two features are strongly related since the odds ratio[53] is equal to $\frac{14\times6}{2\times1} = 42$. As a matter of fact, many studies[54,55] showed that some respondents quickly select the midpoint on the 5-point scale as a dumping ground[56]. Such attitude can be explained in psychological terms: *"choosing a minimally acceptable response as soon as it is found, instead of putting effort to find an optimal response"*[56]. Therefore, we could easily identify "unreliable annotators" that do not produce reliable judgments, by looking at the respondents who fail the control topics. The results of the pilot survey informed our decision to provide the final survey annotators with the same guidelines, but we asked them to evaluate the coherence of topics on a scale from 1 to 4 to discourage annotators from expressing neutral responses.

The final survey was designed to obtain human judgments to be used as ground truth for comparing our method with state-of-the-art coherence measures.

The annotators of the final survey were 222 PhD students from various departments of the University of Palermo; in this way, we employed

highly educated judges with heterogeneous knowledge within the sample. The 222 judges were asked to assess the coherence of 32 topics (30 genuine and 2 artificial topics) on a Google Form[‖].

Table 2 reports the control topics' scores manual assigned by the 222 annotators. Overall, about 90% of the total (202 out of 222 annotators) succeeded in evaluating both control topics. In the case of the highly coherent topic z_{32}, we considered the ratings equal to 4 to "be successful" since a group of words containing only colours should receive the maximum rating. At the same time, we regarded ratings of 1 or 2 as a success for the incoherent coherent topic z_{31}.

Table 2. Control topics' scores assigned by annotators, reliable annotators are highlighted in red.

		Topic z_{32} scores				
		1	**2**	**3**	**4**	Tot
Topic z_{31} scores	**1**	1	1	2	192	196
	2	0	1	4	10	15
	3	0	0	2	4	6
	4	0	2	0	3	5
	Tot	1	4	8	209	222

Fig 5 reports the frequency distributions of the scores assigned by the annotators to the 30 genuine topics, removing the annotators who failed at least one control topic evaluation.

[‖] https://docs.google.com/forms/d/e/1FAIpQLSdoWQsO3MLMcQZDatkCkrSWaThuuj2D-Wm7sR18cy3x8XiRhw/viewform

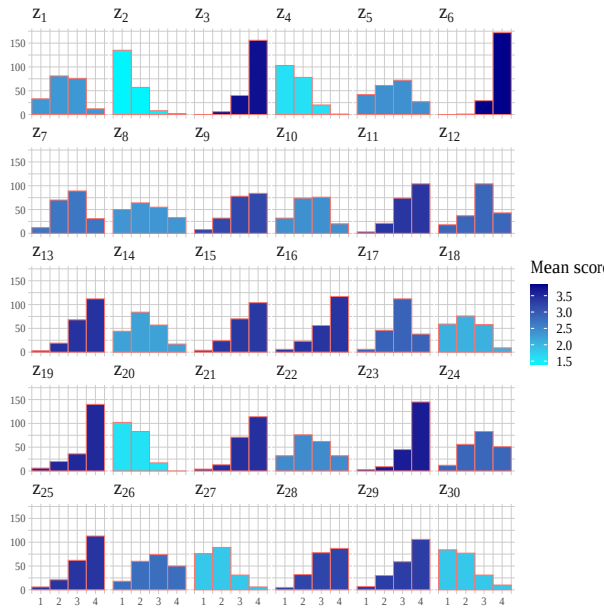


Figure 5. Annotators' coherence evaluations

The final dataset contains: i) the list of the most probable words, ii) the coherence ratings given by evaluators, and iii) the document term matrix used in our study. It is available upon request from the authors.

4.3 Data analysis and results

To compare the effectiveness of the proposed method in replicating human judgment with respect to the other coherence measures proposed in the literature, we collected the results of the survey and re-arranged them in the form of ranking data.

Specifically, a ranking π is a mapping function from the set of topics $\{z_1, \ldots, z_{30}\}$ to the set of ranks $\{1, ..., 30\}$, endowed with the natural ordering of integers; $\pi = (\pi(1), \pi(2), ..., \pi(m))$ where $\pi(z_j)$ is the rank given to topic z_j. In our setting, conditioning to a specific coherence

metric, the topic with the highest coherence score will be ranked 1 and the topic with the lowest coherence score will be ranked 30.
Therefore, we build two matrices:

- the matrix of scores $\mathbf{S}_{30\times 13}$, where the generic s_{ij} element represent the coherence score of the z_j-topic assigned by the i^{th} metric. As regards the last column, i.e. human judgment, the z_j-topic is given the average coherence score assigned by human evaluators. (see Table 3 for a reduced version of the matrix, and Table 6 for the full matrix);
- the matrix of rankings $\mathbf{R}_{30\times 13}$, where the generic r_{ij} element represent the relative rank of the z_j-topic assigned by the i^{th} metric. In this matrix, the estimated topic coherences are compared with each other, in order to establish a preference ordering: from the most coherent to the least coherent topic. (see Table 4 for a reduced version of the matrix, and Table 7 for the full matrix).

Table 3. Coherence scores: the $\mathbf{S}$ matrix

Topic	D_ρ	$\tilde{p}_v$	$\cdots$	HumanJ
z_1	0.076	0.133	$\cdots$	2.332
z_2	0.049	0.084	$\cdots$	1.391
z_3	0.159	0.265	$\cdots$	3.743
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
z_{30}	0.100	0.150	$\cdots$	1.837

Table 4. Ranking coherence scores: the $\mathbf{R}$ matrix

Topic	D_ρ	$\tilde{p}_v$	$\cdots$	HumanJ
z_1	26	26	25	23
z_2	29	29	30	30
z_3	18	18	20	2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
z_{30}	22	23	28	26

To evaluate the correlation between human judgments and the topic quality scores predicted by all the automatic metrics, we use the Emond and Mason's rank correlation coefficient, τ_x [57] (which reduces to Kendal

correlation coefficient τ_b if there are no ties, see[58,59] for an in depth discussion on the correlation measures focusing on the rankings). The higher the τ_x, the better the metric is at measuring topic quality.

In addition, to conforming our comparison procedure to the literature standard, we also computed the Pearson's linear correlation coefficient[10,17] and the Spearman's rank correlation coefficient[11,12,27], see table 8 in the appendix. Although these two measures have been frequently used in the literature, we argue that they are not particularly suitable in this framework. On the one hand, the Pearson's correlation coefficient only considers the linear correlation between two vectors, which is undoubtedly restrictive for our purpose, and its value may be seriously affected by only one outlier[60]. On the other hand, the Spearman rank correlation suffers from the so-called *sensitivity to irrelevant alternatives*, that is: adding extra irrelevant objects to the ranking exercise could change the maximum agreement solution. This issue has been identified by Emond and Mason[61], it is due to the fact that Spearman's correlation estimator treats the ranks as numerical values instead of categorical ordered values. Moreover, Croux and Dehon[60] highlighted that the Spearman rank correlation has a smaller gross error sensitivity (GES) (low robustness) and a greater asymptotic variance (AV) (low efficiency) compared to the Kendall τ_b and τ_x. These features make Spearman coefficient a less preferable estimator from both perspectives.

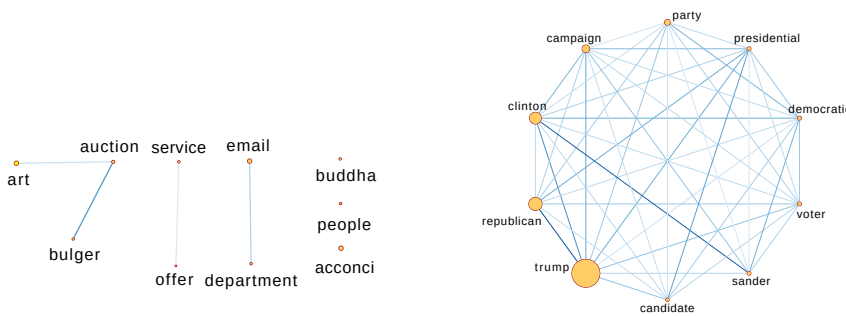
Table 5 reports the τ_x rank correlation between human judgments and all the considered metrics. We compared the correlations obtained either by keeping ("with noise" column of Tab 5) or removing ("without noise" column of Tab 5) the unreliable annotators. The results show that the proposed SVN Coherence measure, based on D_ρ, outperforms all the baselines.

Table 5. Emond and Mason τ_x rank correlation coefficient with human judgments for metrics.

Method	*Correlation with human judgement* τ_x with noise	τ_x without noise
Coh_{SVN}		
J	0.621	0.632
Dc	0.616	0.627
SS	0.616	0.627
FM	0.708	0.714
D_ρ	**0.721**	**0.728**
$\tilde{R}$	0.579	0.586
$\tilde{p}_v$	0.698	0.705
State-of-the-art		
PMI [29]	0.616	0.618
$UMass$ [31]	0.565	0.563
NPMI [10]	**0.685**	**0.687**
CV [17]	0.570	0.572
$tf\text{-}idf$ [16]	0.629	0.636

4.4 Interpretation of the resulting topics

In this section, we report a comparison between Coh_{SVN} and human judgment in evaluating the coherence of some estimated topics. In Fig. 6, topics for which there is high concordance between human judgement and Coh_{SVN} are reported.

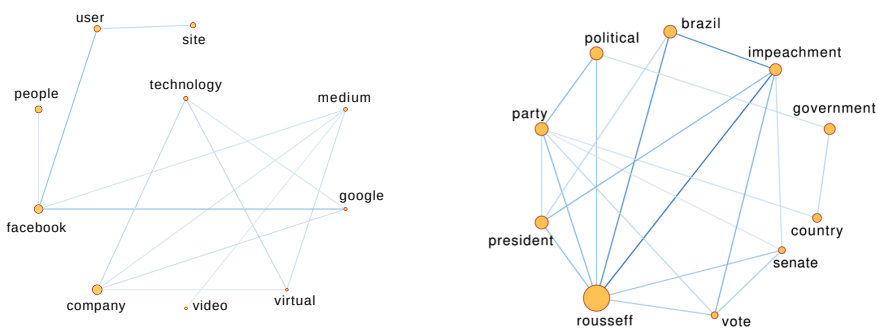


(a) *Topic* z_2 **(b)** *Topic* z_6

Figure 6. SVN representation of Topic z_2 and Topic z_6

In particular, Fig.6(a) represents topic z_6, which is the most coherent topic. It has been assigned an average score equal to 3.84 (first in the rank) by the annotators. Likewise, Coh_{SVN} scores it 0.545, which make it the most coherent in the final ranking. As a matter of fact, topic z_6 can be considered a genuine theme of the domain, i.e., a politically themed topic where all the top words can be associated with US politics. Therefore, the annotators quickly recognized that the words are strongly related, and the co-occurrences in the corpus reflect their solid semantic association.
Topic z_2, in fig. 6(b), is one of the least coherent topics. Annotators rated it with an average score of 1.37 (last position in the ranking). Besides, the topic's Coh_{SVN} score is equal to 0.049, which corresponds to the second-to-last position in the ranking.
The SVN constructed on topic z_2 reveals that the words composing it are mostly unrelated; therefore, there are few statistically validated links.

Fig. 7 report topics whose scores (and, consequently, the ranking) assigned by the annotators are not consistent with our coherence measure.



(a) *Topic* z_3 **(b)** *Topic* z_{28}

Figure 7. SVN representation of Topic z_3 and Topic z_{28}

Topic z_3 (fig. 7(a)) has been positively evaluated by the annotators; the average score is equal to 3.74 (second in the ranking). Instead, Coh_{SVN} places it 18th in the ranking, with a score equal to 0.159
The annotators considered the words in topic z_3 to be related to each other, but the semantic associations detected by humans are not reflected by the co-occurrences in the reference corpus. For example, the words *Facebook* and *company* are not linked in the resulting Statistically Validated Network. This issue could be due to the structure of the corpus used in the analysis. As a matter of fact, the statistical significance of word pairs' co-occurrences can also be validated including external text data sources, such as Wikipedia or Google hits, rather than using only the corpus sentences. Alternatively, one could use paragraphs instead of sentences to count co-occurrences, but if the text is not properly formatted it might prove difficult to identify the paragraphs.
Finally, topic z_{28} is reported in fig. 7(b). The corresponding Coh_{SVN} score is equal to 0.264, the 7th in ranking. While, according to the survey, it has an average score equal to 3.22, and it is 12th in ranking. In this case, the topic is considered to be more coherent by Coh_{SVN} than by humans; however, the discrepancy between the automatic measure and the human judgement is less relevant than in the previous case. Overall, about 20% of the annotators did not recognise a central theme and rated it with a low score (1 or 2). This issue could be since topic z_{28} refers to a specific political event that took place in Brazil between 2015 and 2016. Moreover, it contains "hard-to-interpret" terms such as *Rousseff*, a little-known proper name, and *impeachment*, a technical term referring to the political sphere. Indeed, the evaluation of the topic is more complex than the other ones and requires respondents to carry out in-depth research.

4.5 Summary of main findings

In summary, according to the presented analysis, Coh_{SVN} represents a new topic coherence measure that:

- follows a rigorous statistical model of co-occurrence based on multiple hypotheses testing, while state-of-the-art measures pass over the randomness of co-occurrence;
- ranges between $[0, 1]$, providing a more readable framework for evaluating the coherence of the topics;
- approximates human ratings better than state-of-the-art methods (see Table 5);
- allows the graphical representation and interpretation of the obtained topics through Statistically Validated Networks (SVNs)[19];
- is less sensitive to the text preparation since it considers co-occurrences of word pairs in sentences. Instead, most of the measures proposed in the literature, as summarised in the paper by Röder et al.[17], use a sliding window to calculate the co-occurrences, which makes these methods very sensitive to the preprocessing steps.

5 Conclusions

One of the fundamental challenges in topic detection models is assessing the semantic *coherence* of estimated topics in terms of human interpretability. State-of-the-art coherence measures focus on the marginal probabilities of words and their co-occurrence. However, none of them takes into account the randomness of co-occurrences. In this work, we undertake a rigorous statistical approach based on hypotheses testing to develop a new topic-coherence measure, Coh_{SVN}.

To automatically evaluate how semantically close the top words of the topics are, we represent each topic as a weighted network of its most probable words. The presence of a link between two words indicates that

their co-occurrence in sentences is statistically significant against the null hypothesis of random co-occurrence.

The proposed global measure of coherence, Coh_{SVN}, is derived by considering the number of statistically validated links, the strength of the association between word pairs, and the relative relevance of each word in the topic. To prove the effectiveness of our method, we administered a survey on 222 PhD students from University of Palermo, Italy, and construct a benchmark dataset of human judgements. These judgments were taken as ground truth, and it was shown that the proposed measure reproduces human judgment more closely than the state-of-the-art (Table 5). As for future research, the results reported in this paper suggest to explore the possibility to develop a topic similarity index based on Statistically Validated Networks and including NLP tools, e.g., entity recognition and part-of-speech tagging. Finally, the development of a rigorous statistical method for validating the similarity between two topics could prove beneficial, following the theory of recommendation systems[62], to promote *diversity* in the final ranking of topics. Indeed, the ordered list of topics could be determined by considering both the point-wise quality score (Coh_{SVN}) and the correlations between topics.

References

1. Feldman R, Dagan, I. Knowledge Discovery in Textual Databases (KDT). In *KDD*, vol 95; 1995. p. 112-117.
2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E., Gutierrez, J. & Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv Preprint ArXiv:1707.02919*. 2017.
3. Berkhin, P. A survey of clustering data mining techniques. In *Grouping Multidimensional Data*, 2006; pp. 25-71.
4. McGregor, A., Hall, M., Lorier, P. & Brunskill, J. Flow clustering using machine learning techniques. *International Workshop On Passive And Active Network Measurement*, 2004; pp. 205-214.

5. Wu, W., Xiong, H. & Shekhar, S. Clustering and information retrieval. *Springer Science & Business Media*, 2003.
6. Blei, D., Ng, A. & Jordan, M. Latent dirichlet allocation. *Journal Of Machine Learning Research*, 2003; 3, 993-1022.
7. Blei, D. Probabilistic topic models. *Communications Of The ACM*, 2012; 55, 77-84 .
8. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. Reading tea leaves: How humans interpret topic models. *Advances In Neural Information Processing Systems*, 2009; pp. 288-296.
9. Boyd-Graber, J, Hu, Y, Mimno, D et al. *Applications of topic models*, volume 11. now Publishers Incorporated, 2017.
10. Lau JH, Newman D and Baldwin T. 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*. pp. 530–539.
11. Newman D, Lau JH, Grieser K et al. Automatic evaluation of topic coherence. *In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. pp. 100–108.
12. Aletras N and Stevenson M. Evaluating topic coherence using distributional semantics. *In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. pp. 13–22.
13. Ramrakhiyani N, Pawar S, Hingmire S et al. Measuring topic coherence through optimal word buckets.*In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers. pp. 437–442.
14. Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., Resnik, P.: Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *Advances in Neural Information Processing Systems,* 34, 2021.
15. AlSumait, L., Barbará, D., Gentle, J. & Domeniconi, C. Topic significance ranking of LDA generative models. *Joint European Conference On Machine Learning And Knowledge Discovery In Databases*. Springer, pp. 67–82.
16. Nikolenko SI, Koltcov S and Koltsova O. Topic modelling for qualitative studies.*Journal of Information Science* 2017; 43(1): 88–102.
17. Röder M, Both A and Hinneburg A. Exploring the space of topic coherence measures. *In Proceedings of the eighth ACM international conference on Web search and data mining*. pp. 399–408.
18. Zuo Y, Zhao J and Xu K. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems* 2016; 48(2): 379–398.

19. Tumminello M, Micciche S, Lillo F et al. Statistically validated networks in bipartite complex systems. *PloS one* 2011; 6(3): e17994.
20. Hofmann, T. Probabilistic latent semantic indexing. *Proceedings Of The 22nd Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 50-57, 1999.
21. Blei D and Lafferty J. Correlated topic models. *Advances in neural information processing systems* 2006; 18: 147.
22. Li, W. & McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings Of The 23rd International Conference On Machine Learning*. pp. 577-584, 2006.
23. Dieng, A., Ruiz, F. & Blei, D. The dynamic embedded topic model. *ArXiv Preprint ArXiv:1907.05545*; 2019.
24. Dieng, A., Ruiz, F. & Blei, D. Topic modeling in embedding spaces. *Transactions Of The Association For Computational Linguistics*. **8** pp. 439-453, 2020.
25. Srivastava, A. & Sutton, C. Autoencoding variational inference for topic models. *ArXiv Preprint ArXiv:1703.01488*, 2017.
26. Wallach HM, Murray I, Salakhutdinov R et al. Evaluation methods for topic models. *In Proceedings of the 26th annual international conference on machine learning*. pp. 1105–1112.
27. Morstatter F and Liu H. In search of coherence and consensus: Measuring the interpretability of statistical topics. *Journal of Machine Learning Research* 2018; 18(169): 1–32.
28. Wang L, Wei B and Yuan J. Topic discovery based on lda col model and topic significance re-ranking. *JCP* 2011; 6(8): 1639–1647.
29. Newman D, Karimi S and Cavedon L. External evaluation of topic models. In *Australasian Doc. Comp. Symp.*, 2009. Citeseer.
30. Bouma G. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 2009; : 31–40.
31. Mimno D, Wallach H, Talley E et al. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pp. 262–272.
32. Omar M, On BW, Lee I et al. Lda topics: Representation and evaluation. *Journal of Information Science* 2015; 41(5): 662–675.
33. Salton G and Buckley C. Term-weighting approaches in automatic text retrieval. *Information processing & management* 1988; 24(5): 513–523.
34. Genova VG, Tumminello M, Enea M et al. Student mobility in higher education: Sicilian outflow network and chain migrations. *Electronic Journal of Applied Statistical Analysis* 2019;

12(4): 774–800.

35. Puccio E, Vassallo P, Piilo J et al. Covariance and correlation estimators in bipartite complex systems with a double heterogeneity. *Journal of Statistical Mechanics: Theory and Experiment* 2019; 2019(5): 053404.
36. Kaya B. Hotel recommendation system by bipartite networks and link prediction. *Journal of Information Science* 2020; 46(1): 53–63.
37. Paranyushkin D. Identifying the pathways for meaning circulation using text network analysis. *Nodus Labs* 2011; 26.
38. Miller J. Rg (1981): Simultaneous statistical inference.
39. Arun R, Suresh V, Madhavan CV et al. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 391–402.
40. Krasnov F and Sen A. The number of topics optimization: Clustering approach. *Machine Learning and Knowledge Extraction* 2019; 1(1): 416–426.
41. Sbalchiero S and Eder M. Topic modeling, long texts and the best number of topics. some problems and solutions. *Quality & Quantity* 2020; : 1–14.
42. Chuang J, Gupta S, Manning C et al. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*. PMLR, pp. 612–620.
43. Real R and Vargas JM. The probabilistic basis of jaccard's index of similarity. *Systematic biology* 1996; 45(3): 380–385.
44. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; 26(3): 297–302.
45. Sokal RR, Sneath PHA et al. Principles of numerical taxonomy. *Principles of numerical taxonomy*, 1963.
46. Fowlkes EB and Mallows CL. A method for comparing two hierarchical clusterings. *Journal of the American statistical association* 1983; 78(383): 553–569.
47. Xing L, Paul MJ and Carenini G. Evaluating topic quality with posterior variability. *arXiv preprint arXiv*:190903524 2019.
48. Hong L and Davison BD. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. pp. 80–88.
49. Schütze H, Manning CD and Raghavan P. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.
50. Denny MJ and Spirling A. Text preprocessing for unsupervised learning : Why it matters, when it misleads, and what to do about it. *Political Analysis* 2018; 26(2): 168–189.

51. Bastani K, Namavari H and Shaffer J. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications* 2019; 127: 256–271.
52. Waldherr A, Heyer G, Jaähnichen P et al. Mining big data with computational methods. In *Political Communication in the Online World*. Routledge, 2015. pp. 201–217.
53. Schmidt CO and Kohlmann T. When to use the odds ratio or the relative risk? *International journal of public health* 2008; 53(3): 165.
54. Pornel JB and Saldaña GA. Four common misuses of the likert scale. *Philippine Journal of Social Sciences and Humanities University of the Philippines Visayas* 2013; 18(2): 12–19.
55. Taherdoost H. What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/likert scale. *Hamed Taherdoost* 2019; : 1–10.
56. Chyung SY, Roberts K, Swanson I et al. Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement* 2017; 56(10): 15–23.
57. Emond EJ and Mason DW. A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis* 2002; 11(1): 17–28.
58. Plaia A, Buscemi S and Sciandra M. Consensus measures among preference rankings: a new weighted correlation coefficient for linear and weak orderings. *Journal of Classification*. 2021, 1-22.
59. Albano A and Plaia A. Element weighted Kemeny distance for ranking data. *Electronic Journal of Applied Statistical Analysis*, 14(1) 2021; : 117–145.s.
60. Croux C and Dehon C. Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications* 2010; 19(4): 497–515.
61. Emond EJ and Mason DW. *A new technique for high level decision support*. Department of National Defence Canada, Operational Research Division, 2000.
62. Zhou, Tao, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang, Y. C. : Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 2010; 107(10): 4511-4515.

6 Appendix

Table 6. Coherence scores

	Coh_{SVN}						state-of-the-art						
Topic	J	Dc	SS	FM	D_ρ	$\tilde{R}$	$\tilde{p}_v$	PMI [29]	$UMass$ [31]	NPMI [10]	CV [17]	$tf\text{-}idf$ [16]	HumanJ
z_1	0.006	0.012	0.003	0.137	0.076	0.037	0.133	-2.619	-5.988	-0.119	0.326	-296.42	2.332
z_2	0.004	0.008	0.002	0.089	0.049	0.022	0.084	-5.979	-9.360	-0.272	0.309	-492.41	1.391
z_3	0.010	0.018	0.005	0.291	0.159	0.060	0.265	0.926	-1.965	0.084	0.663	-87.84	3.743
z_4	0.007	0.014	0.004	0.156	0.086	0.042	0.144	-6.258	-9.391	-0.208	0.392	-498.77	1.599
z_5	0.006	0.011	0.003	0.140	0.078	0.037	0.131	-3.533	-6.818	-0.148	0.321	-344.72	2.416
z_6	0.070	0.128	0.037	0.937	0.545	0.422	0.966	1.632	-0.900	0.257	0.899	-5.06	3.847
z_7	0.021	0.039	0.011	0.345	0.194	0.118	0.317	0.864	-1.365	0.127	0.627	-62.09	2.688
z_8	0.009	0.017	0.005	0.228	0.125	0.058	0.219	0.846	-1.826	0.004	0.562	-98.04	2.351
z_9	0.024	0.043	0.013	0.356	0.206	0.148	0.354	-1.721	-5.016	-0.067	0.293	-247.09	3.178
z_{10}	0.018	0.033	0.009	0.277	0.159	0.127	0.281	-1.674	-4.393	-0.102	0.303	-237.55	2.416
z_{11}	0.019	0.037	0.010	0.397	0.223	0.140	0.394	0.977	-1.738	0.063	0.622	-93.95	3.381
z_{12}	0.017	0.032	0.009	0.359	0.201	0.094	0.348	0.804	-1.437	0.046	0.587	-80.95	2.851
z_{13}	0.061	0.111	0.033	0.886	0.506	0.341	0.848	2.437	-1.716	0.365	0.911	-10.61	3.431
z_{14}	0.007	0.013	0.004	0.178	0.098	0.051	0.178	-0.579	-3.907	-0.013	0.484	-190.96	2.233
z_{15}	0.015	0.028	0.007	0.425	0.234	0.085	0.408	0.586	-1.306	0.093	0.590	-79.56	3.356
z_{16}	0.041	0.078	0.021	0.819	0.460	0.271	0.815	1.351	-1.030	0.229	0.845	-34.27	3.406
z_{17}	0.155	0.014	0.505	0.500	0.247	0.048	0.026	-2.040	-4.797	0.007	0.384	-258.17	2.901
z_{18}	0.047	0.002	0.081	0.080	0.008	0.009	0.005	-2.305	-4.582	-0.117	0.234	-277.77	2.084
z_{19}	0.028	0.051	0.015	0.623	0.343	0.152	0.603	1.654	-0.990	0.230	0.867	-32.36	3.535
z_{20}	0.006	0.012	0.003	0.103	0.059	0.050	0.105	-7.400	-11.466	-0.304	0.378	-575.41	1.579
z_{21}	0.023	0.042	0.012	0.419	0.235	0.142	0.396	-0.762	-4.153	0.089	0.579	-191.81	3.460
z_{22}	0.014	0.026	0.007	0.292	0.163	0.085	0.261	-3.339	-6.392	-0.090	0.328	-323.97	2.465
z_{23}	0.023	0.043	0.012	0.445	0.247	0.130	0.418	0.818	-1.308	0.125	0.661	-75.73	3.644
z_{24}	0.012	0.022	0.006	0.224	0.126	0.048	0.184	0.298	-1.236	0.023	0.413	-97.63	2.856
z_{25}	0.028	0.053	0.014	0.600	0.333	0.165	0.564	1.183	-1.350	0.169	0.781	-47.16	3.396
z_{26}	0.021	0.039	0.011	0.356	0.201	0.111	0.351	0.515	-1.683	0.061	0.544	-97.11	2.772
z_{27}	0.007	0.012	0.003	0.132	0.075	0.043	0.135	-7.618	-12.673	-0.285	0.379	-589.28	1.837
z_{28}	0.026	0.047	0.014	0.467	0.264	0.139	0.447	1.287	-1.086	0.155	0.740	-58.45	3.223
z_{29}	0.018	0.034	0.010	0.396	0.219	0.104	0.370	0.825	-1.457	0.127	0.674	-72.54	3.307
z_{30}	0.005	0.010	0.003	0.181	0.100	0.048	0.150	-3.518	-7.015	-0.108	0.312	-346.57	1.837

Table 7. Ranking coherence scores

	Coh_{SVN}							state-of-the-art					
Topic	J	Dc	SS	FM	D_ρ	$\tilde{R}$	$\tilde{p}_v$	PMI [29]	$UMass$ [31]	NPMI [10]	CV [17]	$tf\text{-}idf$ [16]	HumanJ
z_1	25	25	25	26	26	28	26	23	23	25	24	23	23
z_2	30	30	30	29	29	30	29	27	27	28	27	27	30
z_3	20	20	20	18	18	19	18	8	16	12	8	12	2
z_4	22	22	22	24	24	27	24	28	28	27	19	28	28
z_5	27	27	27	25	25	29	27	26	25	26	25	25	20
z_6	1	1	1	1	1	1	1	3	1	2	2	1	1
z_7	11	12	11	16	16	13	16	9	9	8	10	7	18
z_8	21	21	21	20	21	20	20	10	15	18	15	16	22
z_9	8	8	8	15	13	7	13	20	22	20	29	20	13
z_{10}	15	15	15	19	19	12	17	19	19	22	28	19	20
z_{11}	13	13	13	11	11	9	11	7	14	13	11	13	9
z_{12}	16	16	16	13	15	16	15	13	10	15	13	11	16
z_{13}	2	2	2	2	2	2	2	1	13	1	1	2	6
z_{14}	23	23	23	23	23	21	22	17	17	19	17	17	24
z_{15}	17	17	17	9	10	17	9	14	6	10	12	10	10
z_{16}	3	3	3	3	3	3	3	4	3	4	4	4	7
z_{17}	7	6	7	6	6	5	6	21	21	17	20	21	14
z_{18}	29	29	29	30	30	25	30	22	20	24	30	22	25
z_{19}	4	5	4	4	4	6	4	2	2	3	3	3	4
z_{20}	26	26	26	28	28	22	28	29	29	30	22	29	29
z_{21}	10	10	9	10	9	8	10	18	18	11	14	18	5
z_{22}	18	18	18	17	17	18	19	24	24	21	23	24	19
z_{23}	9	9	10	8	8	11	8	12	7	9	9	9	3
z_{24}	19	19	19	21	20	24	21	16	5	16	18	15	15
z_{25}	5	4	5	5	5	4	5	6	8	5	5	5	8
z_{26}	12	11	12	14	14	14	14	15	12	14	16	14	17
z_{27}	24	24	24	27	27	26	25	30	30	29	21	30	26
z_{28}	6	7	6	7	7	10	7	5	4	6	6	6	12
z_{29}	14	14	14	12	12	15	12	11	11	7	7	8	11
z_{30}	28	28	28	22	22	23	23	25	26	23	26	26	26

Table 8. Spearman rank correlation coefficient and Pearson correlation coefficient with human judgments for metrics without noise

	Correlation coefficient without noise	
Method	Spearman	Pearson
J	0.81	0.67
Dc	0.81	0.68
SS	0.81	0.67
FM	0.86	**0.78**
D_ρ	**0.87**	0.77
$\tilde{R}$	0.79	0.66
$\tilde{p}_v$	0.86	0.77
PMI [29]	0.80	0.84
$UMass$ [31]	0.75	0.81
NPMI [10]	**0.88**	**0.87**
CV [17]	0.77	0.76
$tf\text{-}idf$ [16]	0.81	0.85