Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v17n3p545

Beyond human labelling: an automatic topic identification framework for big web data

By Gerli et al.

15 December 2024

Electronic Journal of Applied Statistical Analysis
Vol. 17, Issue 03, December 2024, 545-571
DOI: 10.1285/i20705948v17n3p545

Beyond human labelling: an automatic topic identification framework for big web data

Silvio Gerli[a,b], Roberto Ascari[a], Sonia Migliorati[a], Teresa Cigna[b], and Matteo Borrotti*[a]

[a] *University of Milano-Bicocca, Department of Economics, Management and Statistics (DEMS), Piazza dell'Ateneo Nuovo 1, Milano, 20126, Italy.*
[b] *R & D Department, Sinte Srl, Via Monte Grappa 4, Monza, 20900, Italy.*

15 December 2024

Nowadays, the global amount of written texts grows faster and faster. Since 2011 the number of posts per minute on Facebook increased from 650K to 3M. These unstructured data represent the source of an enormous amount of information that should be extracted by using automatic engines. This can be mainly accomplished by employing Natural Language Processing (NLP), which is a field of Artificial Intelligence devoted to analyzing and understanding human language as it is spoken and written. One common task of NLP is topic identification, related to the recognition of a text's topic(s). Two popular methods for modeling latent topics are latent Dirichlet allocation (LDA) and correlated topic model (CTM). Both assume that each word composing a document is associated with a latent topic, but they differ in the prior distribution assigned to topics, thus showing different pros and cons. In this work, LDA and CTM are tested and compared in a big-data context by analyzing a large set of short documents automatically downloaded from the web by employing a modern crawler. In addition, under the assumption that each document is associated with a single topic, two new methods for the automatic classification of documents according to their real topic are proposed and tested relying on LDA and CTM as (latent) topic model engines. Finally, under the more realistic hypothesis of multiple topics within a document, the two new methods together with some combinations of the two are tested as multi-class classification tools.

*Corresponding authors: matteo.borrotti@unimib.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

keywords: latent Dirichlet allocation, correlated topic model, automatic classification, textual data, topic identification.

1 Introduction

Natural Language Processing (NLP) methods (Manning and Schutze, 1999; Powers and Turk, 1990) are living a very engaging period. The importance of automatic understanding of human language is becoming evident in more and more applications. A few examples of the enormous amount of applicative contexts of NLP are automatic document classification, identification of important contents (*e.g.*, for privacy, classified information, or intellectual property information), automatic categorization of customers' tickets, chatbot questions understanding, and information retrieval from the web.

Many approaches for knowledge extraction from text collections are proposed in the literature. Following the classification proposed by Misuraca and Spano (2020), approaches can be divided into three main classes, *i.e.*, factorial-based approaches, network-based approaches, and probabilistic approaches. The probabilistic approach comprises a family of generative statistical models used for uncovering semantic patterns that reflect the underlying (yet not directly observable) topics within a collection of documents. In the context of NLP, the identification of this(these) underlying topic(s) inside a document, *i.e.*, topic modeling (TM), is particularly useful for discovering the statistical regularities hidden in textual data in supervised/semi-supervised/unsupervised settings (Jelodar et al., 2019). Within TM, the most flexible and widespread probabilistic tools are *latent topic engines*, which allow to detect latent topics associated with each word in a corpus (*i.e.*, a collection) of documents (Blei, 2012).

TM methods have become popular for discovering latent semantic topic structures in small corpora of long documents. However, the massive advancement of communication and information technologies, together with the advent of internet media, information websites, and social media platforms as sources of huge volumes of data, has changed the typical data structure. Indeed, nowadays, *i.e.*, in the big-data era, we often face large corpora of short documents (Murshed et al., 2022). This naturally induces complexity, high variability, and sparsity (Yan et al., 2013; Sridhar, 2015), thus requiring ad hoc strategies. As an example of an ad hoc strategy for handling the problem of sparsity, Yan et al. (2013) introduced a novel approach to modeling topics in short texts, known as the biterm topic model (BTM). In BTM, topics are learned by directly modeling the generation of word co-occurrence patterns, specifically biterms, within the entire corpus.

Differently from the previous approach, we focus on two popular probabilistic TM methods, namely latent Dirichlet allocation (LDA, Blei et al. (2003)) and correlated topic model (CTM, Blei and Lafferty (2007)) because of their acknowledged relevance and widespread employment in the context of TM applications. In particular, LDA was proposed within a Bayesian framework as an enhancement of probabilistic latent semantic analysis (Hofmann, 1999, 2001), and still represents the reference model for TM. CTM is one of the most important proposals alternative to LDA, as it conceives non-negative correlations between topics. Indeed, both methods assume that documents are probability distributions over the topics, whereas topics are defined as probability distributions over the set of unique words composing the corpus. The key difference between the two methods lies in the prior distribution that is assumed for the probability vector of the topic distributions, which results in greater computational tractability as well as ease of interpretation for LDA vs. a more flexible correlation structure between topics for CTM.

The main contribution of the present paper is to propose and validate some new methods to automatically associate the latent unlabeled topics detected by the two above-described TM methods on the one side, with real topics on the other. These two new classification tools are based on an empirical distribution and a word count, respectively. Moreover, some combinations of the two tools are also validated.

More specifically, the threefold goal of the paper is the following. Firstly, LDA and CTM are deployed in a big-data scenario to compare their performance through several widespread indicators. Secondly, assuming that each document within a corpus is associated with a single underlying topic, we introduce and implement two novel methods for the automatic classification of documents into their respective topics. In particular, we resort to LDA and CTM as (latent) topic model engines, and then we apply the two newly proposed methods for topic labelling, *i.e.*, we *baptize* latent topics with a real topic name. Thirdly, we test the proposed classification methods together with some of their combinations as multi-class classification tools under the more realistic assumption that a document is characterized by more than one real topic.

The rest of the paper is organized as follows. Section 1.1 presents the state-of-the-art on automatic labeling methods. Section 2 is devoted to the description of the methodology. In particular, LDA and CTM models (Section 2.1) together with some performance measures (Section 2.2) are described. In Section 3, two new and innovative tools to automatically assign a latent topic to a real topic are illustrated. Section 4 describes the data creation procedure and the resulting dataset, while Section 5 illustrates the final results. Section 6 summarizes some conclusions and hints for future works.

1.1 Automatic topic labelling: literature review

TM produces a collection of latent topics, where each topic is described by a distribution of words. The association of a semantic meaning to these word distributions is not always straightforward. Traditionally, this task is left to human interpretation. However, in the last 15 years, an increasing number of works proposed approaches for automatic topic labelling.

Mei et al. (2007) proposed an unsupervised probabilistic framework to automatically assign a label to a topic model. The authors defined an optimization problem where the final aim was to minimize the Kullback-Leibler divergence between a given topic and the candidate labels and, at the same time, to maximize the mutual information between the two word-distributions. Lau et al. (2010) developed a method for labelling topics based on the top-n terms. The method exploits different ranking mechanisms based on pointwise mutual information and conditional probabilities.

Methods relying on external sources for automatic labelling of topics include the work by Magatti et al. (2009) which derived candidate topic labels for topics induced by LDA using the hierarchy obtained from the Google Directory service, and expanded through the use of the OpenOffice English Thesaurus. Lau et al. (2011) generated label candidates for a topic based on top-ranking topic terms and titles of Wikipedia pages. Then, they built a Support Vector Regression model for ranking the label candidates. Hulpus et al. (2013) developed an automatic topic labelling approach by using a structured data source (DBpedia[1]), and deploying graph centrality measures for generating candidate labels that can characterize the content of a topic.

More recently, Allahyari et al. (2017) proposed a knowledge-based topic model, namely KB-LDA, which integrates the previously mentioned structured data, DBpedia, as a knowledge base for the statistical topic models. Wood et al. (2017) introduced a semisupervised LDA model, Source-LDA. The authors proposed to use labelled knowledge sources representing known potential topics to set the hyper-parameters of the Dirichlet distribution over words. He et al. (2021) introduced a novel two-phase neural embedding framework with a redundancy-aware graph-based ranking process. Furthermore, they provided an up-to-date state-of-the-art analysis.

For a more comprehensive literature review, one can also refer to Misuraca and Spano (2020). The authors give an in-depth description of the process of preparing a collection of documents for quantitative analysis. Moreover, they compare various approaches for automatically extracting information distinguishing the methods into three classes: (*i*) factorial-based approaches, (*ii*) network-based approaches, and (*iii*) probabilistic approaches.

In the present work, we put forward some rules for automatic topic labelling that do not rely on external sources of information. Our proposals, which belong to the class of probabilistic approaches and are based on simple processing of the parameter estimates of LDA and CTM models, succeed in combining automatism in a big-data context and an intuitive interpretation.

2 Latent topic models and performance measures

In this section, we introduce the LDA and CTM models and some model performance indicators.

2.1 Topic models

Let D be the number of documents belonging to a corpus, the d-th document having N_d words ($d = 1, \dots, D$). The set $\mathcal{V}$ of unique words appearing in the corpus has cardinality V, and it is referred to as “vocabulary”.

TM techniques are based on the “bag-of-words” assumption, which implies that word order in a document is irrelevant. The only relevant information in a document is the number of times (*i.e.*, the frequency) each word appears in the document itself. The

[1] `http://dbpedia.org`

"bag-of-words" coincides with an exchangeability assumption for the words within a document.

A further assumption of these approaches is the representation of a document as a probability distribution over a set of K (latent) topics, where a topic is represented by a distribution over words (*i.e.*, with support the vocabulary $\mathcal{V}$). Thus, a document can be depicted as a point $\boldsymbol{\theta}$ in the K-part *topic simplex* $\mathcal{S}^K = \left\{\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\intercal, \theta_k > 0, \sum_{k=1}^K \theta_k = 1\right\}$, whereas a topic corresponds to a vector $\boldsymbol{\phi}$ belonging to the V-part *word simplex* $\mathcal{S}^V = \left\{\boldsymbol{\phi} = (\phi_1, \dots, \phi_V)^\intercal, \phi_v > 0, \sum_{v=1}^V \phi_v = 1\right\}$.

Once the K topic-specific word distributions $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K$ have been generated from a proper distribution, the generative process for the d-th document can be summarized by the following steps:

1. sample $\boldsymbol{\theta}_d$ from a distribution $\mathcal{F}$ defined on $\mathcal{S}^K$;
2. for the n-th word of the document $(n = 1, \dots, N_d)$:
   a) sample a topic $z_{d,n}$ from $Z_{d,n} \sim Categorical(\boldsymbol{\theta}_d)$;
   b) sample a word $w_{d,n}$ from $W_{d,n}|Z_{d,n} = z_{d,n} \sim Categorical(\boldsymbol{\phi}_{z_{d,n}})$.

Both LDA and CTM assume a Dirichlet distribution for the word distribution for the k-th topic:

$$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}), \qquad k = 1, \dots, K, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_V)^\intercal$ and $\beta_v > 0$ for $v = 1, \dots, V$. The main difference between LDA and CTM lies in the distribution for the vector $\boldsymbol{\theta}_d$. Indeed, LDA assumes that

$$\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}), \qquad d = 1, \dots, D, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\intercal$ and $\alpha_k > 0$ for $k = 1, \dots, K$. Differently, CTM assumes that $\boldsymbol{\theta}_d$ follows a logistic-normal distribution (Aitchison, 2003; Blei and Lafferty, 2007), that is the log-ratio transformation $\boldsymbol{\eta}_d = \left(\log\left(\theta_{d1}/\theta_{dK}\right), \dots, \log\left(\theta_{d(K-1)}/\theta_{dK}\right)\right)^\intercal$ is assumed to follow a $(K-1)$-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Therefore, CTM enriches the dependence structure of LDA by including any kind of correlation (*i.e.*, not only negative but also positive) between log-ratio transformed elements of $\boldsymbol{\theta}$. However, this comes at the cost of complicating the interpretation of the dependence structure on the original space. Indeed, there is not a clear relationship between the correlations between log-ratio transformed elements and the original ones. Moreover, and differently from the Dirichlet distribution, the logistic-normal distribution does not possess conjugacy with respect to the categorical distribution, which harms the computational aspects of model inference.

Figures 1 and 2 summarize the LDA and the CTM models through a directed acyclic graph (DAG).

2.2 Typical latent topic indicators

To evaluate the performance of competing models, we resort to two well-established classes of measures, namely coherence and perplexity.

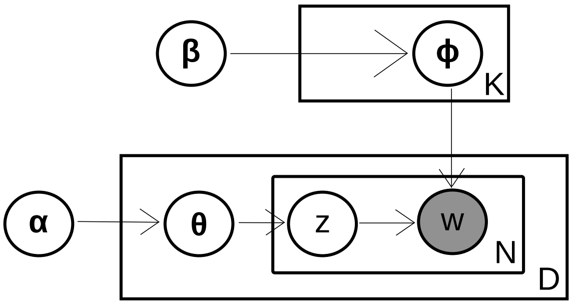

Figure 1: DAG representing the LDA model. Filled nodes represent observed variables.

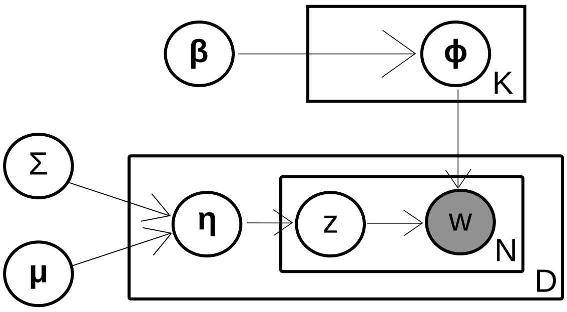

Figure 2: DAG representing the CTM model. Filled nodes represent observed variables.

2.2.1 Coherence

Coherence measures (*e.g.*, see Mimno et al. (2011), Röder et al. (2015), and Syed and Spruit (2017)) have been introduced to assess the intrinsic coherence of the k-th latent topic ($k = 1, \dots, K$) identified by a latent topic model. These measures are strictly connected with the set $\mathcal{I}_k^M$ of the M most probable words in topic k, that are those words associated with the largest estimated elements of $\boldsymbol{\phi}_k$. More precisely, for each pair of words in $\mathcal{I}_k^M$, we compute a "confirmation measure", which is a function depending on the probability $P\left(w_m^{(k)}\right)$ that a document contains the word $w_m^{(k)} \in \mathcal{I}_k^M$ at least once, and the probability $P\left(w_m^{(k)}, w_l^{(k)}\right)$ that a document contains at least once word $w_m^{(k)} \in \mathcal{I}_k^M$ and at least once word $w_l^{(k)} \in \mathcal{I}_k^M$ $(m, l = 1, \dots, M;\ m \neq l)$. Then, a coherence measure for topic k is simply obtained by computing the mean of the confirmation measures over all the pairs of words in $\mathcal{I}_k^M$. Higher values of coherence measures are associated with more interpretable topics.

In particular, we shall use three types of coherence measures based on different confirmation measures. The first considers the pointwise mutual information (PMI) as a confirmation measure:

$$C_{\mathrm{UCI}}^{(k)} = \frac{2}{M \cdot (M-1)} \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \mathrm{PMI}\left(w_m^{(k)}, w_l^{(k)}\right), \tag{3}$$

where

$$\mathrm{PMI}\left(w_m^{(k)}, w_l^{(k)}\right) = \log\left(\frac{P\left(w_m^{(k)}, w_l^{(k)}\right) + \epsilon}{P\left(w_m^{(k)}\right) \cdot P\left(w_l^{(k)}\right)}\right)$$

and ε is a small positive term added to ensure stability of the logarithm function.

A slight modification of PMI is its normalized version (NPMI), which allows defining a second coherence measure:

$$C_{\mathrm{NPMI}}^{(k)} = \frac{2}{M \cdot (M-1)} \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \mathrm{NPMI}\left(w_m^{(k)}, w_l^{(k)}\right), \tag{4}$$

where

$$\mathrm{NPMI}\left(w_m^{(k)}, w_l^{(k)}\right) = \frac{\mathrm{PMI}\left(w_m^{(k)}, w_l^{(k)}\right)}{-\log\left(P\left(w_m^{(k)}, w_l^{(k)}\right) + \varepsilon\right)}.$$

Lastly, we define the topic coherence measure introduced by Mimno et al. (2011):

$$C_{\mathrm{UMass}}^{(k)} = \frac{2}{M \cdot (M-1)} \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{P\left(w_m^{(k)}, w_l^{(k)}\right) + \varepsilon}{P\left(w_l^{(k)}\right)}. \tag{5}$$

Coherence measures can also be aggregated by averaging the same measures over topics (*e.g.*, $C_{\mathrm{UMass}} = \sum_{k=1}^{K} C_{\mathrm{UMass}}^{(k)}/K$).

2.2.2 Perplexity

The perplexity index is a further measure of model performance. In particular, given a new corpus composed of D' unseen documents $\mathcal{C}^T$ (playing the role of a test set), perplexity is computed as

$$\textit{perplexity}\left(\mathcal{C}^T\right) = \exp\left\{-\frac{\sum_{d=1}^{D'} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D'} N_d}\right\}, \tag{6}$$

where $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,N_d})^\intercal$ is the vector of words composing the d-th document, and $p(\mathbf{w}_d)$ denotes the probability assigned by the model to words in document d (i.e., its likelihood). Blei et al. (2003) showed that this measure can be represented as the inverse of the geometric mean of per-word likelihood, thus the larger the likelihood, the smaller the perplexity value. This entails that, in comparing fitted models, the lower the perplexity, the better the model.

3 Automatic labelling of latent topics for documents with a unique real topic

In order to compare LDA and CTM as topic model engines in a classification setting, we assume that a unique "real" topic is associated with each document. Thus, we need an automatic tool able to recognize the real topic of each document among the latent topics detected by the engines.

More precisely, let $\mathcal{C}$ be a corpus of D documents endowed with a set $\mathcal{R} = \{r_1, \dots, r_H\}$ of labels associated with H real topics, and suppose that a latent topic engine has identified $K \geq H$ latent topics. Then, a rule assigning each latent topic to a real topic is needed ("baptism", hereafter). When the assignment is not possible (i.e., when a latent topic does not represent any real topic), the latent topic is baptized as "pseudo-topic". Figure 3 illustrates the rationale of the method in a simple case with $H = 4$ real topics and $K = 16$ latent topics.

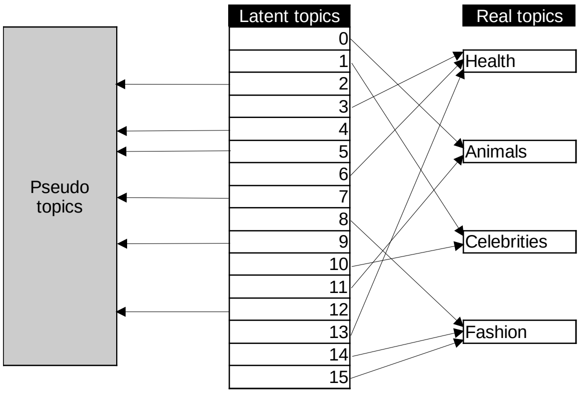

Figure 3: Example of baptism of latent topics as real topics or pseudo-topics.

In the present paper, two methods for baptizing topics are proposed: a "distribution-based" method and a "top words-based" method.

3.1 Distribution-based method

Let $\mathbf{y}$ be the real topics vector, with elements $y_d \in \mathcal{R} = \{r_1, \dots, r_H\}$ representing the real topic of document d, and let $\hat{\boldsymbol{\theta}}_d$ be the score distributions obtained from a latent topic engine (that is, the estimate of the topic-composition of document d, $d = 1, \dots, D$). To assess whether latent topic k should be associated to real topic $r \in \mathcal{R}$, we consider the scores vectors $\tilde{\boldsymbol{\theta}}_k = \left(\hat{\theta}_{1k}, \hat{\theta}_{2k}, \dots, \hat{\theta}_{Dk}\right)^{\intercal}$, $k = 1, \dots, K$, where $\hat{\theta}_{dk}$ is the k-th element of $\hat{\boldsymbol{\theta}}_d$ (i.e., the estimated proportion of topic k in document d).

Then, the distribution-based method sums the elements of $\tilde{\boldsymbol{\theta}}_k$ corresponding to those documents whose real topic is r, that is

$$\tilde{p}_k^r = \sum_{d=1}^{D} \hat{\theta}_{dk} \mathbb{I}\left(y_d = r\right), \qquad r \in \mathcal{R}, \tag{7}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Finally, a probability distribution is computed by normalization:

$$p_k^r = \frac{\tilde{p}_k^r}{\sum\limits_{l \in \mathcal{R}} \tilde{p}_k^l}, \qquad r \in \mathcal{R}. \tag{8}$$

We define the K-dimensional vector $\mathbf{B}$ as the vector with elements b_k representing the real topic that is going to be assigned to latent topic k. To baptize the k-th latent topic, we consider only the largest probability p_k^r and the second largest probability $p_k^{r'}$, and we base the decision on their difference. If this difference is equal to or greater than a given threshold t_d, then latent topic k is baptized as "real topic r" (*i.e.*, the one corresponding to the highest probability) and $b_k = r$, otherwise it is considered as a pseudo-topic and thus discarded. In the case that two latent topics are assigned to the same real topic, then only the one corresponding to the highest probability is considered.

Algorithm 1 summarizes the main steps of the distribution-based method, while Table 1 illustrates it by means of a simple example. Here, a latent topic model has been fitted on a corpus composed of $D = 9$ documents, and $H = 4$ real topics, namely "Health", "Fashion", "Celebrities", and "Animals". The example considers the baptism of latent topic k when $\tilde{\boldsymbol{\theta}}_k = (0.1, 0.8, 0.7, 0.15, 0.1, 0.2, 0.01, 0.9, 0.04)^\intercal$. The latent topic k will be baptized as the real topic "Fashion", since it is the real topic associated with the largest probability (*i.e.*, 0.8), differing from the second largest probability (*i.e.*, 0.13) by more than a threshold t_d equal to 0.2.The latent topic k will be baptized as the real topic "Fashion", since it is the real topic associated with the largest probability (*i.e.*, 0.8), differing from the second largest probability (*i.e.*, 0.13) by more than a threshold t_d equal to 0.2.

Algorithm 1 Pseudo-code: distribution-based method (learning phase)

```
procedure DISTRIBUTION($\mathcal{C}, \mathcal{R}, \tilde{\boldsymbol{\theta}}_k$)
    $\forall r \in \mathcal{R}$ compute $\tilde{p}_k^r = \sum_{d=1}^{D} \hat{\theta}_{dk} \mathbb{I}(y_d = r)$    ▷ $\mathbb{I}(\cdot)$ denotes the indicator function
    Normalization step:
    $\forall r \in \mathcal{R}$ compute $p_k^r = \frac{\tilde{p}_k^r}{\sum_{l \in \mathcal{R}} \tilde{p}_k^l}$
    $\hat{r}^{top1} \leftarrow$ real topic with the highest value $p_k^r$
    $\hat{r}^{top2} \leftarrow$ real topic with the second highest value $p_k^r$
    if $P(\hat{r}^{top1}) - P(\hat{r}^{top2}) > t_d$ then    ▷ $t_d$ is a fixed threshold
        $b_k \leftarrow$ real topic $\hat{r}^{top1}$
    else
        $b_k \leftarrow$ "pseudo-topic"    ▷ The latent topic $k$ is discarded
    end if
end procedure
```

	Real topic			
Documents	**Health**	**Fashion**	**Celebrities**	**Animals**
Doc1 (Health)	0.1			
Doc2 (Fashion)		0.8		
Doc3 (Fashion)		0.7		
Doc4 (Animals)				0.15
Doc5 (Health)	0.1			
Doc6 (Health)	0.2			
Doc7 (Celebrities)			0.01	
Doc8 (Fashion)		0.9		
Doc9 (Celebrities)			0.04	
$\tilde{p}_k^r$ (Equation (7))	0.4	2.4	0.05	0.15
p_k^r (Equation (8))	**0.13**	**0.8**	0.02	0.05

Table 1: Toy example - Baptism of latent topic k by means of the distribution-based method with threshold $t_d = 0.2$.

3.2 Top words-based method

The second method we propose compares the most probable words recovered by a latent topic engine and the most frequent words appearing in a real topic.

Given a corpus $\mathcal{C}$ and a set of real topics $\mathcal{R}$ as in Algorithm 1, for each real topic $r \in \mathcal{R}$ we define the set $\mathcal{T}_5^r$ of top-5 unique words, namely the set of the five most frequent words in documents for which $y_d = r$. The term "unique" means that words in $\mathcal{T}_5^r$ and $\mathcal{T}_5^{r'}$ are selected such that $\mathcal{T}_5^r \cap \mathcal{T}_5^{r'} = \varnothing$ for any $r \neq r'$. We also select the set $\mathcal{I}_k^{10}$ containing the ten most probable words of latent topic k (i.e., the ten words associated to the largest values in $\hat{\boldsymbol{\phi}}_k$, which is an estimate of $\boldsymbol{\phi}_k$, the word distribution for topic k, defined in Section 2.1). Then, we compute the frequency of words in $\mathcal{I}_k^{10}$ which appear in each of the sets $\mathcal{T}_5^r, r \in \mathcal{R}$, and focus on the difference between the largest and the second largest frequencies. If this difference is equal to or greater than a given threshold t_{tw}, then latent topic k is baptized with the real topic r corresponding to the highest frequency, otherwise, it is considered as a pseudo-topic and thus discarded. The method is summarized in Algorithm 2, whereas Table 2 shows a simple example with $t_{tw} = 2$. In the example, the latent topic k will be baptized as the real topic "Health", since it has the highest frequency (*i.e.*, 4) and the difference between this value and the second largest frequency (*i.e.*, 2) is greater than or equal to $t_{tw} = 2$. Then, the result will be stored in position k of the vector $\mathbf{B}$, by assigning $b_k =$ "Health".

Algorithm 2 Pseudo-code: top words-based method (learning phase)

```
procedure TopWords($\mathcal{C}, \mathcal{R}, \mathcal{I}_k^{10}$)
                                        ▷ $\mathcal{I}_k^{10}$, ten most probable words of latent topic $k$
    $\forall r \in \mathcal{R}$ find $\mathcal{T}_5^r$
    Initialize $w$ as a vector of 0 with size equal to the number of real topics
    $r \in \mathcal{R}, w_r \leftarrow$ frequency of words in $\mathcal{I}_k^{10}$ that appear in each of the sets $\mathcal{T}_5^r$
    $\hat{r}^{top1} \leftarrow$ real topic with the highest value $w_r$
    $\hat{r}^{top2} \leftarrow$ real topic with the second highest value $w_r$
    if $P(\hat{r}^{top1}) - P(\hat{r}^{top2}) \geq t_{tw}$ then          ▷ $t_{tw}$ is a fixed threshold
        $b_k \leftarrow$ real topic $\hat{r}^{top1}$
    else
        $b_k \leftarrow$ pseudo-topic                          ▷ The latent topic $k$ is discarded
    end if
end procedure
```

Real topics	**Frequency**	**Top 5 words**
Health	4	**‘patients’, ‘treatment’, ‘visit’**, ‘medicine’, **‘report’**
Animals	0	‘dog’, ‘cow’, ‘cat’, ‘animals’, ‘veterinary’
Celebrities	2	**‘photos’**, ‘gossip’, ‘event’, **‘daughter’**, ‘vip’
Fashion	1	‘fashion’, ‘look’, ‘showroom’, ‘collection’, **‘style’**
Latent topic k		‘patients’, ‘treatment’, ‘visit’, ‘daughter’, ‘car’, ‘report’, ‘photo’, ‘street’, ‘style’, ‘word’

Table 2: Toy example - Baptism of latent topic k by means of the top words-based method with threshold $t_{tw} = 2$.

3.3 Inference and indicators for classification purposes

Both the distribution-based and the top words-based methods rely on estimates of parameters characterizing the underlying LDA and CTM engines described in Section 2.1. LDA model can be fitted by either a variational inference approach, as originally proposed by Blei et al. (2003), or by a fully collapsed Gibbs sampling (Griffiths and Steyvers, 2004). The latter approach can be improved by partitioning the data across separate processors and performing inference in parallel, as suggested by Newman et al. (2009). The same holds for the CTM, whose parameters can be estimated by a mean-field variational inference algorithm (Blei and Lafferty, 2007) or by an efficient Gibbs sampling algorithm as described in Mimno et al. (2008). Moreover, inference for unseen documents is based on techniques from discriminative text classification as proposed by Yao et al. (2009). The main idea of their approach is to move the collapsed Gibbs for additional iterations on an “updated” corpus including also the unseen documents. At the end of these additional iterations, we obtain an estimate for the vector of topic proportions of the new documents, namely $\hat{\boldsymbol{\theta}}_{new}$. Additional details on the implementation of these techniques can be found in the cited works.

We now illustrate how we can build a classifier to classify unseen documents to their unique real topic. Let us consider a new document, its estimated vector $\hat{\boldsymbol{\theta}}_{new}$, and the vector $\mathbf{B}$ obtained by one of the two baptizing methods. For ease of explanation, we consider a toy example considering a latent topic engine fitted considering $K = 8$ latent topics. A baptizing method produced the vector $\mathbf{B}$ = ("Health", "pseudo-topic", "Health", "Fashion", "Celebrities", "pseudo-topic", "Fashion", "Animals"), whereas the predicted vector of topic proportion resulted equal to $\hat{\boldsymbol{\theta}}_{new} = (0.2, 0.05, 0.15, 0.02, 0.12, 0.18, 0.1, 0.18)^\intercal$. Classification is performed by summing the elements in $\hat{\boldsymbol{\theta}}_{new}$ having with the same label in $\mathbf{B}$. Thus, the topic-label "Health" leads to a total proportion equal to $\hat{\theta}_{new,1} + \hat{\theta}_{new,3} = 0.2 + 0.15 = 0.35$, whereas the other labels lead to $\hat{\theta}_{new,4} + \hat{\theta}_{new,7} = 0.12$ ("Fashion"), $\hat{\theta}_{new,5} = 0.12$ ("Celebrities"), $\hat{\theta}_{new,8} = 0.18$ ("Animals"), and $\hat{\theta}_{new,2} + \hat{\theta}_{new,6} = 0.23$ ("pseudo-topic"). Given these totals, if we have to classify the new document to a unique real topic, the topic with the largest total proportion will be chosen for the final classification. In the above example, the new document is going to be classified as a "Health"-related document. We can extend this rule to consider the case we have to classify a "mix" (*i.e.*, a document characterized by more than one real topic). Let us assume that the mix is composed of two real topics, then the two topics with the largest total proportion will be selected. In the toy example proposed in Table 3, the second largest total proportion is associated with the "pseudo-topic" label. When such a situation occurs, the topic with the largest total proportion among those remaining is selected as the second real topic. In our case, the new document is going to be classified with the mix ("Health", "Animals").

b_k	Health	pseudo	Health	Fashion	Celebr.	pseudo	Fashion	Animals
$\hat{\theta}_{new,k}$	0.2	0.05	0.15	0.02	0.12	0.18	0.1	0.18

Table 3: Toy example - Classification of a new document having its predicted vector of topic proportions $\hat{\boldsymbol{\theta}}_{new}$ and the vector of baptized topics $\mathbf{B}$.

In the following, we shall compare the distribution-based and the top words-based methods using standard classification performance indicators such as the *accuracy* (*i.e.*, the number of documents correctly assigned to the corresponding real topic over the total number of documents). Since we are also interested in the *true positive rate* and the *positive predictive value*, *recall* and *precision* are considered too. More precisely, for each real topic $r \in \mathcal{R}$, *recall* of topic r is defined as the ratio between the number of documents correctly assigned to topic r (called the true positive value of real topic r, TP_r) and the number P_r of documents with actual real topic r. The *precision* of topic r is defined as the ratio between TP_r and the number P_r^* of documents that are baptized as real topic r. *Recall* and *precision* can be used to compute the *F1-score*. In the binary classification setting, *F1-score* is defined as the harmonic mean of the two measures. Since we test our methods on a problem characterized by H real topics, we also use the so-called *macro F1-score* (Sokolova and Lapalme, 2009), which is computed by considering the average *precision* and the average *recall* over the H real topics.

4 Data collection and dataset creation

High dimensionality is a crucial issue while treating textual data. Classical approaches should be tested in complex scenarios and, more importantly, proper tools for the automatic identification of topics are needed to tackle these scenarios. For this reason, a novel high-dimensional dataset has been created. The set of data should have some peculiarities: (i) a large number of documents should be considered, (ii) a real topic should be associated with each document, and (iii) mixed real topics (*i.e.*, more than one real topic) should be involved in the definition of the ground truth. The latter request allows testing the proposed methods as multi-class classification tools. Indeed, mixed real topics are often encountered in real data applications. Thus, testing LDA and CTM together with the proposed automatic classification methods in the context of mixed real topics is an important issue.

To achieve this result, we resorted to a web crawler. The latter is a software able to browse a website (in our case, a mono-thematic portal) and download text from articles following all internal links of the website. Thus, two types of recorded information are provided: (a) the article texts and (b) the real topic (or mixed real topics) of the texts.

The identification of a unique real topic for a text is ensured by selecting specialized websites (*e.g.*, a website about animals). The mixed real topics are identified considering the structure of the website. Indeed, portals are organised in subsections, typically sub-homepages, on which articles are related both to the main topic of the portal and to the topic of the subsection (*e.g.*, “Health” - “Animals”). Thus, all texts downloaded from a subsection are labelled with mixed real topics. In this work, we consider $H = 4$ real topics (*i.e.*, specialized websites) and two mixed real topics (specialized websites with subsections), as shown in Table 4.

(Mixed) Real Topics	**N. of docs**
Health	6.868
Animals	3.914
Celebrities	14.095
Fashion	14.064
Health - Animals	279
Celebrities - Fashion	227

Table 4: Dataset structure: Real and mixed real topics (column 1) and number of downloaded documents for each (mixed) real topic (column 2).

The dataset creation process begins by compiling a list of URLs for each topic, sourced from mono-thematic websites or specific subsections of websites. Subsequently, a crawler is deployed to download textual content associated with each URL. This process involves several steps. Firstly, the landing page's content (*i.e.*, the page's main text) for each URL is downloaded using the `essence`[2] library, designed for extracting web page information. To collect documents referring to the chosen topic, all links within the landing page containing the original URL are identified. For example, if the landing page is `https://www.nytimes.com/section/world`, then only URLs starting with that string (*e.g.*, `https://www.nytimes.com/section/world/europe`) are considered. For each selected URL, the crawler downloads the content and identifies all the links within the page. This recursive process continues until no additional links are found. The downloaded documents, as well as their true topic, are then gathered together. The set of downloaded documents has a dimension equal to 91.342. A common pre-processing phase has been performed on the corpus. Firstly, we conducted a filtering step, removing all duplicates (*i.e.*, those accidentally downloaded twice or more) and empty documents. In this step, non-Italian documents were also excluded. Indeed, considering more than one language would affect the structure of latent topics since we could obtain language-specific topics (*e.g.*, topics for Italian and non-Italian documents) instead of context-specific topics. A second step in the pre-processing procedure involved the words (*i.e.*, tokens) in the corpus. More specifically, we converted all words to lowercase and removed all the numbers, double spaces, symbolic tokens (*e.g.*, @ and +), punctuation, and single-character words. Additionally, we removed all the sentences that are not related to the documents (*e.g.*, advertisement or cookie consent messages), URLs and hyperlinks, and Italian stopwords. For the list of Italian stopwords, we resorted to the `nltk` Python library. Finally, we removed all the documents that had less than 200 words. The final corpus $\mathcal{C}$ contains $D = 39.447$ documents and is composed of a vocabulary of $V = 228.060$ unique words. Figure 4 shows the texts' length before and after the pre-processing phase for each (mixed) real topic. The type-token ratio (TTR) of the final corpus, namely the ratio between the dimension of the vocabulary and the total number of words composing the corpus, is equal to 0.025, whereas the percentage of hapax legomenon (*i.e.*, words appearing once in the corpus) is equal to 42.4%. These two measures suggest that the corpus is predominantly composed of 57.6% of vocabulary terms, which are frequently repeated.

5 Experimental setting and results

In this Section, we present the results of three experiments on the dataset described in the previous section. In particular, the corpus $\mathcal{C}$ is used to accomplish the present work's three main objectives. In the first instance (Section 5.1), the whole corpus $\mathcal{C}$ (all six rows of Table 4) is used to test LDA and CTM in a big-data scenario. Secondly, the texts labeled with a unique real topic (first four rows of Table 4) are used for testing the distribution-based and top words-based methods as classifiers with LDA and CTM being

[2] `https://github.com/essence/essence`

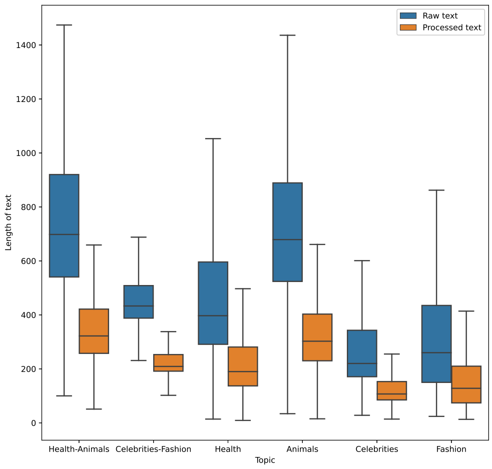

Figure 4: Box plot representation of the length of texts before (blue) and after (orange) the pre-processing activity.

used as latent topic engines (Section 5.2). Lastly, the texts associated with mixed real topics (last two rows of Table 4) are used to conduct a preliminary study to understand the ability of the two novel classification methods as multi-classifiers, still with LDA and CTM being used as latent topic engines (Section 5.3).

In all experiments, the corpus $\mathcal{C}$ is divided into two disjoint parts, namely a training set (composed of 80% of documents), and a test set (the remaining 20% of documents). The random splitting is performed by stratifying over real topics. Once the training and test sets are obtained, we performed an additional pre-processing step, aimed at increasing the classifier's accuracy. Once we obtained the training and test sets, for each topic, we listed the 200 "most frequent words" and, to ensure purer topics, we removed from the corpus those most frequent words that appear in at least three out of four lists.

All experiments are performed in Python, by using the `tomotopy`[3] library to estimate both LDA and CTM models. All elaborations have been carried out with an Intel Core i7 with RAM 16 GB. The code, hyper-parameter settings for LDA and CTM, and additional materials are available at `https://github.com/matteoborrotti/automaticlabeling`.

5.1 LDA and CTM in a big-data scenario

The performance of LDA and CTM in our big-data scenario is evaluated by resorting to the perplexity and coherence metrics presented in Section 2.2. We consider a number K of latent topics ranging from 2 to 16, thus also including the "true" number of topics we considered in generating the corpus, i.e., the value 4.

[3] `https://bab2min.github.io/tomotopy/`

To select the best value of K from a predictive perspective, we rely on the perplexity measure computed on the unseen corpus, that is our test set. Figure 5 shows the perplexity of the two models for several values of K. The perplexity decreases as the number of topics increases, thus suggesting to prefer the value $K = 16$ for both the LDA and CTM models. Nonetheless, from Figure 5 it emerges that LDA performs far better than CTM for any value of K.

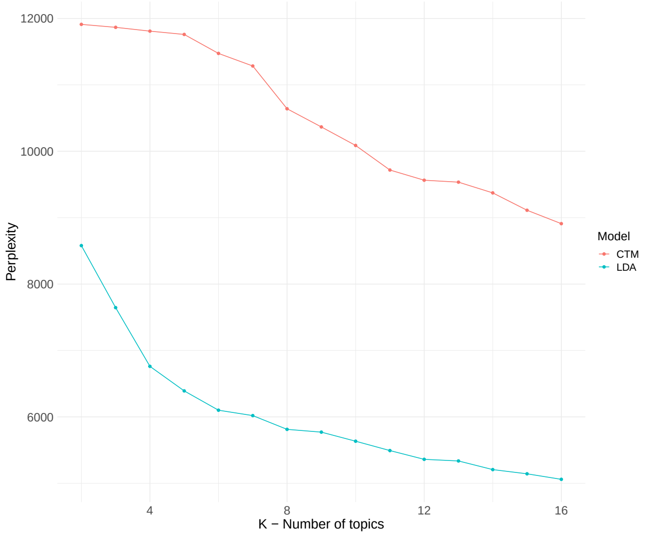

Figure 5: Big-data scenario. Perplexity of LDA and CTM as K increases.

The quality of the recovered topics can be inspected by looking at the most probable words for each model (e.g., inspecting the word clouds as in Figure 6), and accordingly assigning a label to each latent topic in the light of these words. Table 11 in Appendix 1 reports the 20 most probable words for the LDA model. It emerges that LDA identifies distinct topics that human judgment can hardly categorize in the four considered real topics "Health", "Animals", "Celebrities", and "Fashion". In particular, LDA recognizes some new topics (e.g., topic 2 is related to "Music", topics 3 and 6 deal with different aspects of "Beauty" routines) and splits some real topics into subtopics (e.g., topics 4, 15, and 16 split the Animal category into "Dogs", "Non-pets", and "Cats", respectively).

Contrarily, by inspecting the most probable words for the CTM model with $K = 16$ latent topics (Table 12 in Appendix 1), it is evident that there is no such clear semantic homogeneity in the recovered topics.

Interestingly, a value of $K = 16$ much larger than 4 (i.e., our ground truth) seems to perform better with both the LDA and the CTM models. Indeed, this result is coherent with the conclusion of some authors affirming that the perplexity measure often selects a number of topics that is too large (Jingxian and Yong, 2021; Sbalchiero and Eder, 2020). For this reason, we prefer to inspect also the coherence results for four values of K. For each $K \in \{4, 8, 12, 16\}$, we compute the coherence measures (applying averages and standard deviations over topics of Equations (3), (4), and (5)) of the two models

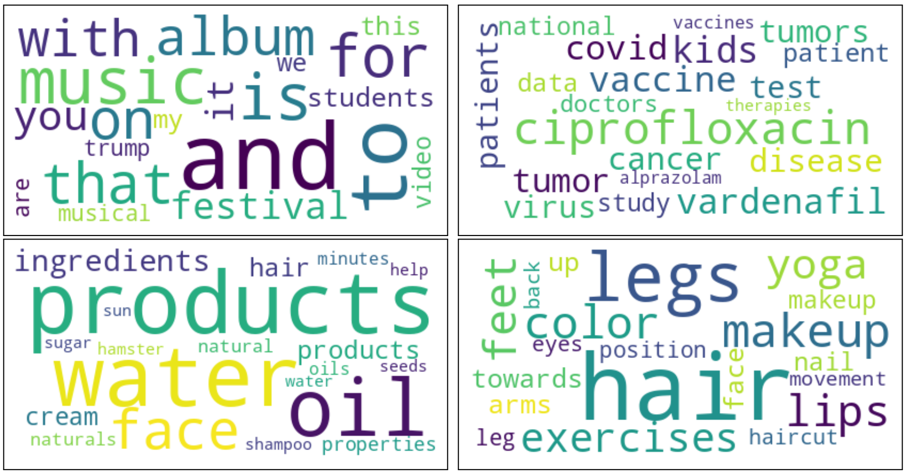

Figure 6: Word clouds representing four LDA latent topics. Detected latent topics refer to “Music” (topic 2), “Health” (topic 7), and different aspects of “Beauty” (topics 3 and 6).

on the training corpus, by considering the $M = 10$ most probable words within each topic. Table 5 summarizes the coherence results by model and number of topics K. By inspecting Table 5, one can note that topics generated by the LDA model with $K = 4$ are considered the most coherent by all the coherence measures, since they are characterized by the largest mean and smallest standard deviation. Contrarily, the number of topics maximizing the coherence measures in the CTM varies between 12 and 16.

In conclusion, the LDA model seems to reliably identify topics both with $K = 4$ and $K = 16$, with a different granularity level (e.g., with more latent topics, it can discover “new” relevant topics in the corpus), whereas the CTM points to a larger number of topics that are not semantically homogeneous.

Model	**K**	$C_{\textbf{UCI}}$	$C_{\textbf{NMPI}}$	$C_{\textbf{UMass}}$
LDA	4	**0.76** (0.36)	**0.11** (0.03)	**-1.41** (0.4)
	8	0.45 (1.13)	0.1 (0.05)	-1.99 (1.12)
	12	0.68 (1.01)	**0.11** (0.05)	-1.99 (0.89)
	16	0.29 (1.62)	0.1 (0.09)	-2.29 (1.27)
CTM	4	-3.54 (1.86)	-0.17 (0.07)	-4.4 (1.79)
	8	-1.07 (2.14)	-0.01 (0.12)	-2.85 (1.56)
	12	**-0.49** (1.78)	**0.03** (0.11)	**-2.17** (0.85)
	16	-0.69 (1.81)	**0.03** (0.1)	-2.39 (0.99)

Table 5: Big-data scenario. Mean and standard deviation (in parenthesis) of coherence measures stratified by model and number of topics. Best values are reported in bold.

5.2 Automatic identification of unique real topics

The distribution-based method (hereafter, D) and the top words-based method (T) described in Sections 3.1 and 3.2 are tested as automatic classification techniques for the identification of real topics. In this context, LDA and CTM are used as latent topic engines for both methods. Furthermore, in order to label a text, we also test two different combinations of the two methods. More precisely, if the two methods strictly agree on the topic identification, then that topic is assigned to the text, otherwise, no topic is assigned (*i.e.*, the topic is labelled as pseudo-topic). From now on, this combination is referred to as the $D \bigwedge T$ approach. A further combination is also considered, which assigns a topic whenever the two methods agree on the topic (as in the $D \bigwedge T$ approach), but also when one method identifies a topic while the other method detects a pseudo-topic. In all other cases (*i.e.*, the two methods disagree on the topic or they both identify a pseudo-topic), the topic is labeled as pseudo-topic. This approach is denoted by $D \bigvee T$.

The performance of the D and T methods depends on different parameter choices. Among these choices, one of the most relevant is the selection of the thresholds for defining whether a text is associated with a real topic or a pseudo-topic. For this reason, a sensitivity study is performed to ascertain the influence of parameters t_d and t_{tw} defined in Algorithms 1 and 2. More precisely, we compare the classification performances considering $t_d \in \{0, 0.1, 0.2, 0.3\}$ for the D method, $t_{tw} \in \{1, 2, 3\}$ for the T method, and $K \in \{4, 8, 12, 16\}$.

		Distribution				**Top-words**		
		t_d				**t_{tw}**		
	K	**0**	**0.1**	**0.2**	**0.3**	**1**	**2**	**3**
LDA	4	0.72	0.77	0.77	0.77	0.77	0.77	0.77
	8	**0.85**	**0.85**	0.84	0.78	0.76	0.73	0.73
	12	0.82	**0.85**	**0.85**	**0.85**	0.77	0.73	0.49
	16	0.84	0.84	**0.85**	**0.85**	**0.79**	0.76	0.46
CTM	4	0.43	0.00	0.00	0.00	0.10	0.10	0.00
	8	0.46	0.36	0.00	0.00	0.37	0.48	0.36
	12	0.49	0.40	0.54	0.00	0.45	0.44	0.44
	16	0.58	0.59	**0.60**	0.58	**0.62**	0.45	0.37

Table 6: Global accuracy for D and T methods. Best values are reported in bold.

Table 6 shows the global accuracy of D and T methods. It emerges that considering the LDA as the latent topic engine helps the whole set of methods to get larger (i.e., better) values than the ones obtained by adopting the CTM model, for any values of t_d and t_{tw}. The performance of the D method together with LDA seems to be independent of either K and t_d, with the only exception $K = 4$. If we consider the T method combined with LDA, a careful selection of t_{tw} should be done. A larger value of t_{tw} leads to smaller values of global accuracy. This is even more evident if we also consider a larger value of K. More generally, the farther K from the ground truth, the more strict the threshold t_{tw} should be.

		Distribution				**Top-words**		
		$\mathbf{t_d}$				$\mathbf{t_{tw}}$		
	K	**0**	**0.10**	**0.20**	**0.30**	**1**	**2**	**3**
LDA	4	0.57	0.58	0.58	0.58	0.75	0.75	0.75
	8	**0.84**	**0.84**	0.81	0.75	0.75	0.73	0.73
	12	0.79	0.83	0.82	0.83	0.76	0.76	0.50
	16	0.83	0.83	**0.84**	**0.84**	**0.78**	0.76	0.47
CTM	4	0.25	0.00	0.00	0.00	0.08	0.05	0.00
	8	0.32	0.13	0.00	0.00	0.38	0.40	0.13
	12	0.40	0.29	0.31	0.00	0.45	0.28	0.28
	16	0.43	0.47	**0.48**	0.34	**0.66**	0.43	0.36

Table 7: Macro *F1-score* for D and T methods. Best values are reported in bold.

		$\mathbf{D \bigvee T}$				$\mathbf{D \bigwedge T}$			
		$\mathbf{t_d}$				$\mathbf{t_d}$			
K	$\mathbf{t_{tw}}$	**0**	**0.1**	**0.2**	**0.3**	**0**	**0.1**	**0.2**	**0.3**
4	1	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**
	2	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**
	3	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**
8	1	**0.85**	**0.85**	0.84	0.78	**0.76**	**0.76**	**0.76**	**0.76**
	2	**0.85**	**0.85**	0.84	0.78	0.73	0.73	0.73	0.73
	3	**0.85**	**0.85**	0.84	0.78	0.73	0.73	0.73	0.73
12	1	**0.85**	0.84	0.83	0.73	**0.77**	**0.80**	**0.80**	**0.80**
	2	0.82	**0.85**	**0.85**	**0.85**	0.83	0.73	0.73	0.73
	3	0.82	**0.85**	**0.85**	**0.85**	0.49	0.49	0.49	0.49
16	1	0.84	0.84	**0.85**	**0.85**	**0.79**	**0.79**	**0.79**	**0.79**
	2	0.84	0.84	**0.85**	**0.85**	0.76	0.76	0.76	0.76
	3	0.84	0.84	**0.85**	**0.85**	0.46	0.46	0.46	0.46

Table 8: Global accuracy for $D \bigvee T$and $D \bigwedge T$ methods for the LDA model. Best values are reported in bold.

The values of macro *F1-score* are also reported in Table 7. Focusing on the LDA model as the latent topic engine, we notice that the D method gets larger values of macro *F1-score* with $K = 8$ and $K = 16$. If we focus on the T method, the situation is slightly different. In fact, larger values are reached only if we consider $K = 16$. Globally, the D method obtains higher values of macro *F1-score* with respect to the T method.

Given these results, we focus on the comparison of $D \bigvee T$ and $D \bigwedge T$ only considering LDA as the latent topics engine. From the results in Table 8, the combination of the D and T methods does not lead to better performance. $D \bigvee T$ reaches the same results as the D method, therefore there is no advantage in considering the use of the two methods together. Moreover, the results lead us to assume that the D is the method ruling the final decision of the $\bigvee$ logical operator.

5.3 Test on the identification of mixed real topics

Since a document may be described by more than one topic, there is an urgent need to develop suitable approaches that recognize mixed real topics. For this purpose, distribution-based and T methods are tested as multi-class classification tools. Similarly to Section 5.2, also here LDA and CTM are used as latent topic engines in both methods. $D \bigwedge T$ and $D \bigvee T$ are also tested. Again, all the combinations between $K \in \{4, 8, 12, 16\}$, $t_d \in \{0, 0.1, 0.2, 0.3\}$, and $t_{tw} \in \{1, 2, 3\}$ are tested. As reported in Table 4, we consider two mixed topics, that are "Health" - "Animals" and "Celebrities" - "Fashion".

		Distribution				**Top-words**		
		$\mathbf{t_d}$				$\mathbf{t_{tw}}$		
	K	**0**	**0.1**	**0.2**	**0.3**	**1**	**2**	**3**
LDA	4	0.44	0.44	0.44	0.44	**0.80**	**0.80**	**0.80**
	8	0.95	0.95	0.96	0.64	0.60	0.61	0.61
	12	0.92	0.97	**0.98**	0.95	0.55	0.61	0.17
	16	0.91	0.91	0.92	0.92	0.52	0.54	0.13
CTM	4	**0.45**	0.00	0.00	0.00	0.55	0.00	0.00
	8	**0.45**	0.00	0.00	0.00	0.70	0.41	0.00
	12	**0.45**	**0.45**	**0.45**	0.00	0.48	0.00	0.00
	16	**0.45**	**0.45**	**0.45**	**0.45**	**0.94**	0.36	0.47

Table 9: Global accuracy for D and T methods. In this case, two mixed topics are considered. Best values are reported in bold.

Table 9 summarizes the main results in terms of global accuracy. Similar to the previous case, the performance of the D method with LDA as the topic engine is independent of K and t_d, except for $K = 4$. Increasing the number K of latent topics affects the performance of the T method (with LDA). When we set a value of K higher than 4, results start to deteriorate. Also in this case, the use of CTM as the topic engine is not advisable.

		$\mathbf{D \bigvee T}$				$\mathbf{D \bigwedge T}$			
		$\mathbf{t_d}$				$\mathbf{t_d}$			
K	$\mathbf{t_{tw}}$	**0**	**0.1**	**0.2**	**0.3**	**0**	**0.1**	**0.2**	**0.3**
4	1	0.44	**0.80**	**0.80**	**0.80**	**0.44**	**0.44**	**0.44**	**0.44**
	2	0.44	**0.80**	**0.80**	**0.80**	**0.44**	**0.44**	**0.44**	**0.44**
	3	0.44	**0.80**	**0.80**	**0.80**	**0.44**	**0.44**	**0.44**	**0.44**
8	1	0.95	0.95	**0.96**	0.64	0.60	0.60	0.60	0.60
	2	0.95	0.95	**0.96**	0.64	**0.61**	**0.61**	**0.61**	**0.61**
	3	0.95	0.95	**0.96**	0.64	**0.61**	**0.61**	**0.61**	**0.61**
12	1	0.97	0.93	0.93	0.89	**0.61**	**0.61**	**0.61**	**0.61**
	2	0.92	0.97	**0.98**	0.95	**0.61**	**0.61**	**0.61**	**0.61**
	3	0.92	0.97	**0.98**	0.95	0.17	0.17	0.17	0.17
16	1	0.91	0.91	**0.92**	**0.92**	0.52	0.52	0.52	0.52
	2	0.91	0.91	**0.92**	**0.92**	**0.54**	**0.54**	**0.54**	**0.54**
	3	0.91	0.91	**0.92**	**0.92**	0.13	0.13	0.13	0.13

Table 10: Global accuracy for $D \bigvee T$and $D \bigwedge T$ methods for the LDA model. In this case, two mixed topics are considered. Best values are reported in bold.

As in Section 5.2, we consider the two approaches $D \bigvee T$ and $D \bigwedge T$ only with LDA as the latent topic engine. Table 10 reports the values of global accuracy. Differently from the previous case, the combination of the D and T methods leads to slightly better performance than the ones obtained by using only one of the two methods. In conclusion, it seems that $D \bigvee T$ outperforms $D \bigwedge T$ and a larger value of K is preferable for classifying documents characterized by a mix of real topics.

6 Conclusions and future work

In this work, two popular latent topic engines (*i.e.*, LDA and CTM) are compared in a big-data scenario characterized by a large number of documents ($\approx$ 39.500). The massive corpus is built with an automatic big-data text collector (*i.e.*, a web crawler) able to search for thousands of texts and associate a real topic to each one of them.

Additionally, an innovative set of methods to associate latent unlabelled topics with real topics is introduced. D method, T method, and two combinations of them ($D \bigwedge T$ and $D \bigvee T$) are proposed as automatic identification tools for real topics. LDA and CTM are used as latent topic engines in all mentioned methods. All approaches are evaluated as classification tools, as well as for their ability to adapt to modern large corpora of short documents.

The comparison between LDA and CTM as latent topic models in a big-data scenario highlights LDA as the best-performing engine. Moreover, also when considering LDA and CTM as latent topic engines for the automatic identification of real topics, LDA outperforms CTM in combination with any of the proposed methods for labelling texts. The D method and the $D \bigvee T$ are the two best approaches for assigning a real topic to a text (the largest values in almost all performance indicators). For all approaches, the most difficult task is the identification of mixed real topics.

In future work, we intend to increase the number of documents, thus enlarging the size of the corpus $\mathcal{C}$. In addition, we plan to consider a higher number of real topics and mixed real topics as ground truth. Another relevant issue to be tackled is the hyper-parameter choice. In the present work, standard settings were resorted to, but a deeper analysis of the hyper-parameter tuning phase could be fruitful. Finally, a larger set of latent topic models could be considered as latent topic engines.

Supplementary information

Supplementary materials, including also all the word clouds, are available at `https://github.com/matteoborrotti/automaticlabeling`.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

7 Appendix

7.1 Appendix 1: tables of Section 5.1

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
	me	and	products	dog	cases	hair	ciprofloxacin	project
	said	to	water	dogs	swab	legs	kids	fashion
	says	is	oil	animals	diet	makeup	vardenafil	vogue
	friends	music	face	pedigree	deaths	color	vaccine	brand
	say	for	ingredients	cats	food	feet	covid	job
	have	that	products	puppy	virus	lips	tumor	milan
	want	on	hair	pedigrees	foods	yoga	cancer	design
	father	album	cream	need	none	exercises	disease	accross
	maybe	with	properties	owner	eat	towards	tumors	city
	that	you	natural	rabbit	phase	arms	virus	designer
	singer	festival	naturals	names	meat	face	patients	fashion
	instagram	it	minutes	animal	protein	up	test	future
	love	students	oils	size	positives	nail	patient	research
	social	we	help	nature	fats	makeup	data	new
	mother	my	hamster	company	quantity	position	study	culture
	written	this	seeds	shepherd	number	leg	national	director
	person	video	shampoo	play	region	haircut	doctors	beauty
	story	musical	sugar	kids	beginning	eyes	alprazolam	space
	no	trump	water	fear	deaths	movement	vaccines	exhibition
	you	are	sun	hello	check	back	therapies	projects
Real Topic	Celebrities	—	—	Animals	Health	—	Health	Fashion
Label	Social Media	Music	Beauty (Products)	Dogs	COVID	Beauty (Body)	Diseases	Fashion

	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
	star	fashion	photo	vip	patients	sleep	species	cat
	cinema	look	daughter	natalia	mg	lansoprazole	animals	dog
	red	portfolio	rome	episode	treatment	blood	fishes	cats
	carpet	style	kaia	mastrota	must	pain	insects	vet
	series	clothes	lex	big_brother	dose	stress	males	fur
	main_character	jeans	enne	brother	paragraph	disease	specimen	horse
	kate	fashion	mum	giulia	effects	oxygen	birds	kitty
	harry	spring	social	men	data	cells	females	might
	queen	clothes	sanremo	photo	administration	symptom	animal	animal
	jennifer	winter	cindy	rodriguez	reaction	issue	turtle	food
	meghan	pants	milan	d'urso	drug	capable	female	teeth
	oscar	brand	birthday	share	renal	factor	male	symptoms
	new	shoes	couple	belen	therapy	brain	big	dogs
	venice	runway	beautiful	tempation	risk	level	turtles	feline
	hollywood	summer	tv	island	drugs	activity	live	paw
	york	maison	chiara	francesco	supervisory_report	physical	mammal	animal
	director	chanael	francesco	balivo	concomitant	cause	little	diseases
	prince	wear	laura	caterina	adverse	system	hunt	puppies
	lady	jacket	son	famous	side	disease	size	issues
	plot	accessories	marco	read	use	wealth	bees	animal
Real Topic	Celebrities	Fashion	Celebrities	Celebrities	Health	Health	Animals	Animals
Label	Gossip	Fashion show	VIP	Reality	Adverse Events	Health	Animals (non-pet)	Cats

Table 11: Section 5.1: big-data scenario. 20 most probable words for each of the $K = 16$ latent topics detected by the LDA model. Please note that words are translated from the Italian language.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
	fashion	story	cat	kaia	audience	dog	patients	equal
	fashion	francesco	dogs	cindy	must	hair	treatment	dog
	brand	that	cats	know	species	dogs	mg	share
	designer	ilary	dogs	news	degree	cat	data	clarithromycin
	project	says	animals	both	meat	hair	paragraph	system
	vogue	mara	symptoms	friend	check	vet	drug	nifedipine
	style	emma	species	smartphone	products	animals	drugs	report
	costanza	francesca	swab	gregoraci	none	vitamin	reaction	general
	hair	hand	food	free	can	diet	therapy	information
	caracciolo	book	must	crawford	reported	animal	dose	effects
	collection	millions	species	hours	effects	quantity	administration	animals
	designer	blasi	renal	gossip	ramipril	foods	effects	doctor
	gucci	hourses	cases	work	effect	benefits	cases	grater
	vieri	travel	puppies	app	size	fruit	side	kids
	boots	marrone	diseases	that_is	sustainable	product	infections	must
	new	all	pet	some	safety	ears	adverse	specialist
	model	little	animal	reality	oil	paw	concomitant	check
	loves	past	kitty	android	prescription	fish	ciprofloxacin	study
	couture	told	animal	download	oil	diabete	eo	data
	haircut	big	test	best	massage	protein	doctor	use
Real Topic	Fashion	VIP	Animals (pet)				Drugs/Desease	
Label	Beauty	Gossip	Salute Animale					

	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
	tail	temptation	natalia	men	look	vip	photo	evening
	virus	colors	mastrota	belen	collection	enne	video	heart
	horse	and	issues	rodriguez	fashion	son	instagram	back
	disease	legs	article	marco	clothes	lex	marriage	daughter
	nails	to	BigBrother	stefano	jeans	social	york	christmas
	vaccine	fabric	episode	d'urso	suit	couple	cinema	week
	fundamental	effect	tells	andrea	spring	wife	birthday	friends
	respect	euro	giorgio	luca	black	mum	star	milan
	birds	position	read	me	pants	husband	september	alessanfra
	towards	island	caterina	brother	accessories	show	main_character	seems
	eyes	high	balivo	name	style	fan	alessia	photo
	age	yoga	giulia	said	red	children	Sunday	few
	need	michael	face	michelle	clothes	maria	daughter	job
	death	leg	sky	also	shoes	elisabetta	tv	evaluation
	risk	neck	repeated	gemma	white	rome	beautiful	sara
	lips	point	birth	federica	models	fabrizio	evening	girlfriends
	cells	oxygen	seventh	say	colors	corona	model	have
	animal	shirt	immediately	crisis	pair	new	super	come
	come	suit	accepts	throne	pink	red	milan	arrived
	covid	simple	navigation	given	color	actress	party	known
Real Topic								
Label		Clothes		VIP				

Table 12: Section 5.1: big-data scenario. 20 most probable words for each of the $K = 16$ latent topics detected by the CTM model. Please note that words are translated from the Italian language.

References

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data.* The Blackburn Press, London, second edition.

Allahyari, M., Pouriyeh, S., Kochut, K., and Arabnia, H. R. (2017). A knowledge-based topic modeling approach for automatic topic labeling. *International Journal of Advanced Computer Science and Applications*, 8(9).

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55:77–84.

Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(SUPPL. 1):5228 – 5235.

He, D., Ren, Y., Khattak, A. M., Liu, X., Tao, S., and Gao, W. (2021). Automatic topic labeling using graph-based pre-trained neural embedding. *Neurocomputing*, 463:596–608.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177 – 196.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 465–474, New York, NY, USA. Association for Computing Machinery.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211.

Jingxian, G. and Yong, Q. (2021). Selection of the optimal number of topics for lda topic model – taaking patent policy analysis as an example. *Entropy*, 23:1301.

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA. Association for Computational Linguistics.

Lau, J. H., Newman, D., Karimi, S., and Baldwin, T. (2010). Best topic word selection for topic labelling. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, page 605–613, USA. Association for Computational Linguistics.

Magatti, D., Calegari, S., Ciucci, D., and Stella, F. (2009). Automatic labeling of topics. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1227–1232.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 490–499, New York, NY, USA. Association for Computing Machinery.

Mimno, D., Wallach, H., and McCallum, A. (2008). Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, volume 61.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 262–272.

Misuraca, M. and Spano, M. (2020). Unsupervised analytic strategies to explore large document collections. In *Text Analytics, Studies in Classification, Data Analysis, and Knowledge Organization*, pages 17–28.

Murshed, B., Mallappa, S., Abawajy, J., Saif, M., Al-ariki, H., and Abdulwahab, H. (2022). Short text topic modelling approaches in the context of big data: tanonomy, survey, and analysis. *Artificial Intelligence Review.*

Newman, D., Asuncion, A., Smyth, P., and Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(62):1801–1828.

Powers, D. and Turk, C. (1990). *Machine Learning of Natural Language.* Springer-Verlag, New York.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Sbalchiero, S. and Eder, M. (2020). Topic modeling, long texts and the best number of topics. some problems and solutions. *Quality and Quantity*, 54:1095–1108.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437.

Sridhar, V. (2015). Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 192–200.

Syed, S. and Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174.

Wood, J., Tan, P., Wang, W., and Arnold, C. (2017). Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 411–422.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Yao, L., Mimno, D., and McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 937–946.