A Fuzzy Topic Modeling Approach to legal corpora

Antonio Calcagnì, Arjuna Tuzzi

Abstract This study investigates the application of Fuzzy Latent Semantic Analysis (fLSA) in analyzing expenditure chapters within legal texts, using Italy's budget law 178/2020 as a case study. Faced with challenges in legal studies, such as specialized language and heterogeneity, fLSA combines Latent Semantic Analysis (LSA) with dimensionality reduction and soft clustering. Results show comparable performance with the widely used Latent Dirichlet Allocation (LDA) in identifying coherent expenditure chapters, with fLSA showing a tendency to retrieve more distinctive and exclusive topics.

Key words: topic modeling, fuzzy modeling, lda, topic coherence

1 Introduction

Topic models are nowadays one of the most relevant techniques for uncovering latent structures in large textual datasets, spanning various domains such as social sciences [9], digital humanities [7], scientometrics [8], and legal studies [6]. Particularly challenging is their application in legal studies, where specialized language and heterogeneity pose difficulties [3, 4]. This complexity is amplified when dealing with short legal corpora, as in the case of budget laws, which serve as critical guides for a country's finances. In this contribution, we adopts a fuzzy topic modeling approach (fLSA) to analyze the Italian budget law 178/2020. This technique combines the strengths of Latent Semantic Analysis (LSA) with dimensionality reduction and soft clustering to handle sparsity and high-dimensionality challenges in word tokens

Antonio Calcagnì,
University of Padova, e-mail: antonio.calcagni@unipd.it
GNCS, National Institute of Advanced Mathematics (INdAM)

Arjuna Tuzzi,
University of Padova, e-mail: arjuna.tuzzi@unipd.it

and features. The focus is on retrieving -which is quite challenging in legal corpora- as well as assessing the capability of fLSA in detecting highly coherent expenditure chapters (i.e., high-quality topics), especially in comparison with widely used techniques such as Latent Dirichlet Allocation (LDA). All the materials like algorithms and datasets used throughout this contribution are available to download at `https://github.com/antcalcagni/flsa_legalCorpora`.

2 Fuzzy Latent Semantic Analysis

In the analysis of text data, a corpus contains n documents of N word-tokens. The i-th document is represented as a vector of N_i words $\mathbf{w}_i = \{w_1, \ldots, w_{N_i}\}$ such that $\sum_{i=1}^{n} N_i = N$. To represent the joint information of documents and lexicon, the document-term matrix $\mathbf{X}_{n\times J}$ is usually constructed (J is the number of word-types), which constitutes the input for the topic analysis. Unlike other LDA-based methods [1], the fLSA algorithm computes the basic topic modeling matrices - namely, $\mathbf{P}_{D|T}$ (probability of documents per topic) and $\mathbf{P}_{W|T}$ (probability of words per topic) - by means of a deterministic algorithm that uses both SVD and fuzzy c-means in its iterative steps [5]. Given a (possibly locally-globally weighted) document-term matrix $\mathbf{X}_{n\times J}$, the following computations are then required for K topics:

$$\mathbf{P}_{D|T_{(n\times K)}} = \left(\Xi_{(n\times K)} \circ \mathbf{p}_{D_{n\times 1}} \mathbf{1}_K^T\right)\ \mathrm{diag}\left(\mathbf{1}_n^T (\Xi_{(n\times K)} \circ \mathbf{p}_{D_{n\times 1}}) \mathbf{1}_K^T\right)^{-1} \tag{1}$$

$$\mathbf{P}_{W|T_{(J\times K)}} = (\mathrm{diag}(\mathbf{1}_n^T \mathbf{X}_{n\times J})^{-1} \mathbf{X}_{n\times J}^T)\ \mathbf{P}_{D|T_{n\times K}} \tag{2}$$

where $\mathbf{p}_D = (\mathbf{1}_n^T \mathbf{X})(\mathbf{I}\, s)^{-1}$, with $s = (\mathbf{1}_n^T \mathbf{X} \mathbf{1}_K)$, is the vector of document probability, Ξ is the matrix of fuzzy membership degrees computed via fuzzy c-means algorithm on the matrix $\mathbf{U}_{n\times K}$ of the first K left singular vectors of $\mathbf{X}$, $\mathrm{diag} : \mathbf{R}^M \to \mathbf{R}^{M\times M}$ is the usual linear operator, whereas $\circ$ denotes the Hadamard product. Put this way, the fLSA generalizes the standard LSA by using the fuzzy c-means (or any other fuzzy clustering technique such as fuzzy k-medoids) to find the unobserved topics through the set of documents. Although it misses a consistent stochastic framework to compute $\mathbf{P}_{D|T}$ and $\mathbf{P}_{W|T}$, the technique has shown optimal results in terms of classification, document clustering, and document redundancy, if compared to other commonly used methods for topic modeling [5].

3 Application: The case of Italian budget law 178/2020

Data and preprocessing

The corpus being analyzed consists of $n = 1162$ clauses referring to the Italian budget law 178/2020.[1] The entire corpus conveys $N = 52223$ word-tokens and $J = 6207$ word-types, which highlights a good level of redundancy (TTR = 11.80%, hapax = 31.20%). The minimum and maximum lengths are 3.00 and 648.00 word-tokens, respectively, with a Q3-Q1 range of 35 word-tokens. The raw corpus has been preprocessed to achieve a standardized format. Hence, all uppercase letters have been converted to lowercase, punctuation marks and symbols have been removed, and a first subset of multiword expressions (e.g., *partita iva*) have been identified and transformed (e.g., *partita_iva*) by matching with a predefined list. Additionally, to reduce unnecessary word redundancy, all words referring to other laws (e.g., *d.l. 30 giugno 1998*) have been considered a single word (e.g., *dl_30_6_1998*), while a second subset of multiword expressions have been identified and created by considering 5-to-2-grams with highest pointwise mutual information (PMI). With the same aim, commonly used modal verbs (e.g., *essere*, *avere*, *dovere*, *potere*) as well as words with less than nine occurrences have been filtered out. As a result, the final corpus has $N = 32878$ word-tokens and $J = 1392$ word-types, with TTR = 6.90%. Finally, the document-term-matrix has been computed and the common term-frequency/inverse-document-frequency (TF-IDF) schema has been adopted to locally/globally weight the ensuing frequency matrix.

Methods and procedures

The performance of fLSA method to retrieve highly coherent topics has been assessed against the LDA method in a repeated 10-fold cross validation schema (with 25 random repetitions). Two commonly used coherence metrics, namely UMass-like (i.e., `mean_logratio`) and UCI-like (i.e., `mean_pmi`) metrics, as implemented in the `R` library `text2vec`, have been chosen for the between-method comparison. The number K of topics has been varied from 2 to 45.

Results

Figure 1 shows the coherence metrics averaged over the 10 folds for both methods. Although the criteria indicate different optimal numbers of topics for fLSA and LDA — specifically, $K^\dagger \in \{11, 12\}$ in the first case and $K^\dagger \in \{7, 14\}$ in the second — both highlight a slightly higher coherence for LDA compared to fLSA.[2]

[1] In this context, each document represents a single clause, as the budget law is structured based on the use of a single article.

[2] For the analysis, the upper bound of $K^\dagger$ has been selected for both methods.

Interestingly, the methods differ in the topic construction, with fLSA producing a less uncertain marginal topic distribution than LSA (Figure 2). Moreover, fLSA detects fewer exclusive topics than LDA, resulting in a higher level of distinctiveness. Based on Frex top words[3], the topics produced by both methods span from labor policies to health policies related to COVID-19, covering social welfare as well. However, if we restrict the focus only to those topics referring to covid-19 pandemic, we notice that fLSA identifies only two ($k \in \{5, 11\}$), which have the largest probability to occur and are highly exclusive. By contrast, LDA identifies five topics of this type ($k \in \{3, 5, 7, 11, 14\}$), which are hard to distinguish from the others. Figure 3 shows the the topic distances in an MDS projection for both methods. Interestingly, the solution provided by fLSA suggests that topics can be aggregated into three clusters (blue dashed rectangles), while no clear hierarchical structure emerges for the LDA method. To see whether the fLSA-based analysis would benefit from a three topics solution, the algorithm has been run by setting $K = 3$. The new solution conveys topics with higher average coherence compared to the case where $K = 12$. These topics are almost uniformly distributed over the clauses ($p_{\text{K=1}} = 0.415$, $p_{\text{K=2}} = 0.361$, $p_{\text{K=3}} = 0.223$) but exhibit different levels of exclusivity ($\text{ER}_{\text{K=1}} = 0.238$, $\text{ER}_{\text{K=2}} = 0.578$, $\text{ER}_{\text{K=3}} = 0.182$). Figure 4 depicts the FREX-based top words for the three topics alongside the word-topic proportion (colored bars). Unlike the previous solution, the latter shows a coarse classification of expenditure chapters into three general macro-areas of economic activity.

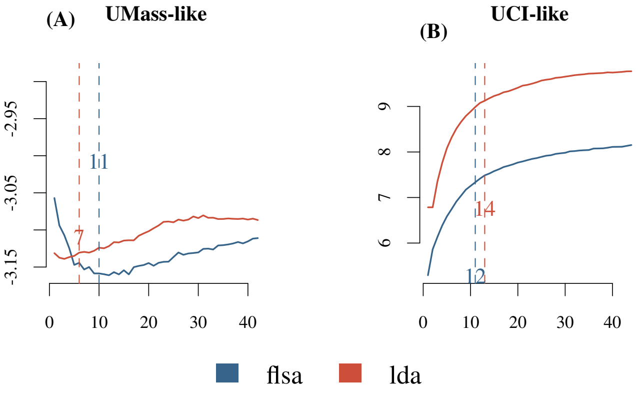

Fig. 1 Average Coherence Scores for fLSA (blue curve), and LDA (red curve) as a function of the number of topics. Note that dashed vertical lines indicate the elbow of the curves, and the numbers indicate the corresponding elbow points (i.e., the optimal number of topics). All metrics have been computed using the top thirty topic words.

[3] FREX-based top words for each topic and both methods are available in the repository indicated in Section 1.

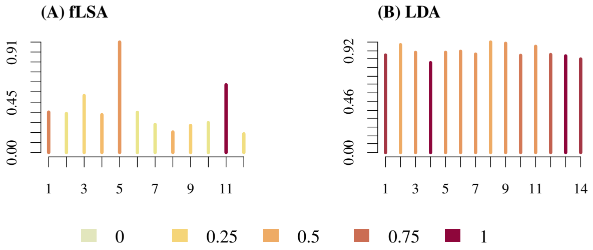

Fig. 2 Marginal topic distributions (bars) alongside the exclusive term ratio (gradient color). Note that frequencies and exclusivity index have been normalized for the sake of comparison.

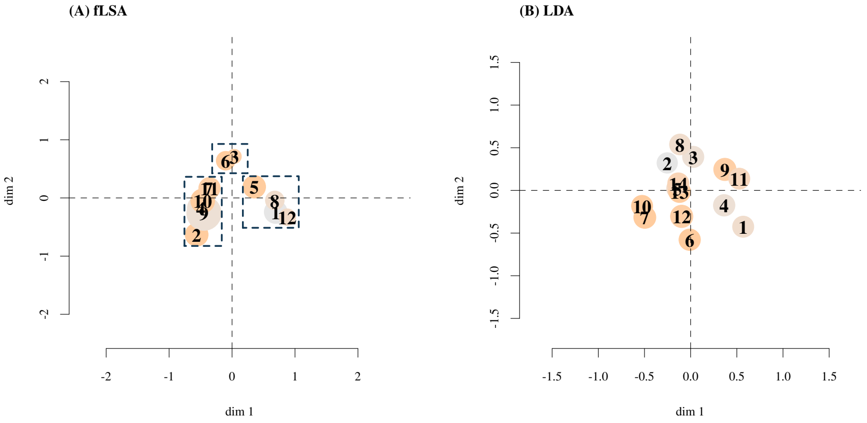

Fig. 3 Inter-topic distance via MDS-based plot for both fLSA and LDA methods. Note that each circle represents a topic, the circle radius represents the marginal probability of the topic (in log scale), the circle color represent the UCI-like coherence (the stronger the saturation of orange, the higher the coherence), whereas the blue rectangles indicate topic clusters extracted via Ward-based hierarchical clustering on the matrix of inter-topic distances.

4 Conclusions

In this contribution, we have explored the feasibility of applying topic modeling to specialized legal texts, particularly in the analysis of expenditure chapters within budget laws. Specifically, Fuzzy Latent Semantic Analysis (fLSA) has been employed on Italy's budget law 178/2020. Our focus has been on assessing the effectiveness of fLSA in identifying coherent expenditure chapters compared to the commonly used Latent Dirichlet Allocation (LDA). The results indicate that fLSA has been capable of retrieving coherent and informative topics to the same extent as LDA. However, they differ in terms of marginal probability and exclusivity, with

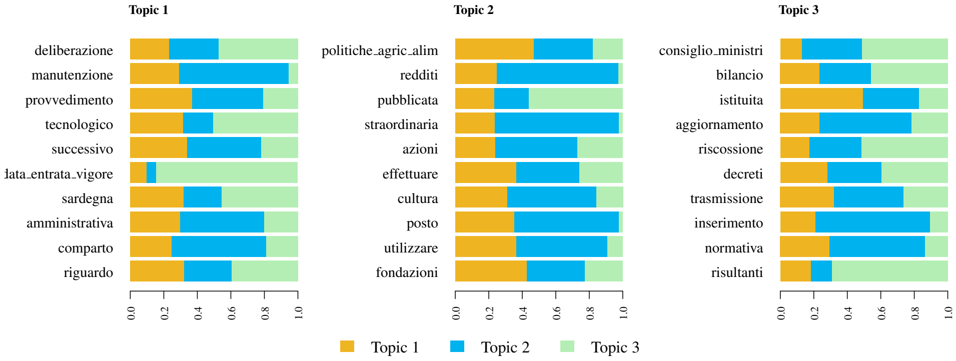

Fig. 4 Highest Frex topic words alongside the proportion of occurrence for each estimated topic (fLSA).

topics tending to cluster when projected onto a lower MDS-based space. Further investigations are needed to evaluate the extent to which fLSA differs from more specialized methods like the Correlated Topic Model (CTM). Similarly, additional studies would be needed to assess how fuzzy membership degrees Ξ can be integrated to compute $\mathbf{P}_{D|T}$. While the current implementation has used a linear aggregator, other solutions can been explored, such as those based on generalized Bayes rules [2].

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
2. Coletti, G., Gervasi, O., Tasso, S., Vantaggi, B.: Generalized bayesian inference in a fuzzy context: From theory to a virtual reality application. Computational Statistics & Data Analysis **56**(4), 967–980 (2012)
3. Cortelazzo, M.: La lingua delle leggi italiane. In: M.E. Piemontese, et al. (eds.) Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile, pp. 110–122. Carocci (2023)
4. Garavelli, B.M.: Le parole e la giustizia. Einaudi (2001)
5. Karami, A., Gangopadhyay, A., Zhou, B., Kharrazi, H.: Fuzzy approach topic discovery in health and medical corpora. International Journal of Fuzzy Systems **20**, 1334–1345 (2018)
6. Luz De Araujo, P.H., De Campos, T.: Topic modelling brazilian supreme court lawsuits. In: Legal Knowledge and Information Systems, pp. 113–122. IOS Press (2020)
7. Schröter, J., Du, K.: Validating topic modeling as a method of analyzing sujet and theme. Journal of Computational Literary Studies **1**(1) (2022)
8. Tuzzi, A., et al.: Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Springer (2018)
9. Valdez, D., Pickett, A.C., Goodson, P.: Topic modeling: latent semantic analysis for the social sciences. Social Science Quarterly **99**(5), 1665–1679 (2018)