What do we learn by applying multiple methods in topic detection? An empirical analysis on a large online dataset about mobility electrification

Che cosa impariamo applicando diversi metodi per identificare gli argomenti di un corpus? Analisi empirica su un grande insieme di dati online sull'elettrificazione della mobilità

Fabrizio Alboni, Margherita Russo and Pasquale Pavone[*]

Abstract Identifying the topics covered in a corpus is one of the central issues in automatic text analysis. The objective of our paper is to contribute to the comparative analysis of different methods. In particular, we compare the results obtained through the use of the most common methods for topic identification, applied to the same corpus. The analysis is performed on a large original textual database created from an e-mobility newsletter. To compare the results between the methods, we refer to two criteria. First of all, the semantic consistency of the different models is evaluated by applying the UMass score and Pointwise mutual information. Secondly, the degree of association between the topics identified by the different models is processed using a heat-map and Cramer's V.

Abstract L'identificazione degli argomenti trattati in un corpus è uno dei temi centrali dell'analisi automatica dei testi. Obiettivo del nostro articolo è contribuire all'analisi comparata di diversi metodi. In particolare, confrontiamo i risultati ottenuti attraverso l'uso dei metodi più comuni per l'identificazione di argomenti, applicati allo stesso corpus. L'analisi viene effettuata su un ampio database testuale originale creato a partire da una newsletter sulla mobilità elettrica. Per confrontare i risultati tra i metodi, facciamo riferimento a due criteri. In primo luogo, la coerenza semantica dei vari modelli è valutata applicando il punteggio UMass e il Pointwise mutual information. In secondo luogo, il grado di associazione tra gli argomenti identificati dai diversi modelli viene elaborato con una heat-map e con la V di Cramer.

Key words: topic detection, text mining, Cramer's V, coherence indexes, electric mobility

[*] Fabrizio Alboni, Università degli studi di Modena e Reggio Emilia, fabrizio.alboni@unimore.it
Margherita Russo, Università degli studi di Modena e Reggio Emilia; margherita.russo@unimore.it:
Pasquale Pavone, Università degli studi di Modena e Reggio Emilia, pasquale.pavone@unimore.it:

Introduction

The continuous proliferation and availability of digitized textual information has led, especially over the last two decades, to an increase in demand - in both academia and industry - for systems and algorithms capable of extracting information of interest from unstructured, semi-structured and fully structured textual data. This availability of data makes it possible, on the one hand, to carry out qualitative analyses of document collections in all research contexts and, on the other hand, to develop Apps with the most diverse objectives in the context of everyday life. Research activities in the field of text analysis have developed rapidly: many of Text Mining's approaches [1, 3, 4, 7, 20] effectively combine linguistic resources, computational methods and statistical techniques for the analysis of texts, representing a highly interdisciplinary field. In general, these processes do not involve only the training of the models, but also require numerous additional procedures, pre-processing of texts, transformation and reduction of the dimensionality of the data being analysed.

Among the many objectives that can be defined within a text analysis, of particular importance are the clustering techniques of documents on the basis of their similarity in terms of content, and more in detail the identification of the topics covered in the collection of documents [2, 11, 13]. In this regard, one of the most used methods in this context is topic modelling which, starting from a first work by Blei et al. [5] was developed by Griffiths et al. [8] introducing the Latent Dirichlet Allocation (LDA) as a generative model for identifying topics within a corpus. This method is sometimes overused within any type of context without the necessary adaptation of the analysis strategy to the characteristics of the corpus. As alternatives to LDA, multiple methodologies have been formulated for the exploration of topics, such as: Latent Semantic Analysis (LSA) [6]; the Reinert method [15, 16]; Non-negative Matrix Factorization (NMF) [9]; Correspondence and cluster analysis [10].

The goal of our paper is contributing at the broad debate on text analysis, as it is summarized by Lebart [11]. He focuses on the comparison of different techniques (NMF, LDA, Correspondence Analysis and Clustering) applied on a given middle size and homogeneous corpus, i.e. Shakespeare’s 154 Sonnets.

In our paper we rely on a large original textual database created on a newsletter– issued daily by electrive.com on electric mobility. Russo et al. [18] have already analysed that data set (for the period 21th August 2018-15th September 2021) to identify the emerging topics in a domain of rapid transformation of the automotive industry. Entities disambiguation techniques, topic detection based on Correspondence and cluster analysis have been already commented by Russo et al. [18] who identify eight main classes of topics and 24 subtopics. In this paper, we update the database (with news item until 8th March 2022) and address the exploration of some clustering techniques.

In his analysis, Lebart concludes that the various methods "concur on the same topics […] despite the amazing variety of their theoretical backgrounds", and he underlines that results depend on "on various parameters and options", and that "exploratory or descriptive tools… have been essential to visualize the complexity of the process and to assess the obtained results" [13, p. 11].

In our paper, we refer to the expert classification (directly provided by the newsletter editors) and to three alternative methods for clustering/topic detection, based, respectively, on probabilistic, cluster-based and factorial methods: LDA; Reinert method; Correspondence analysis to select the most relevant factors explaining variability within the corpus, on which a hierarchical cluster analysis is applied. The rational for our choice is argued in the paper together with a discussion of the pros and cons of the various clustering techniques. Ad hoc visual tools for the comparative analysis have been created by using Tableau. Analytical methods to compare the results refer to the hierarchical cluster analysis as a benchmark.

Automatic analysis enables speed, consistency and reproducibility, and produces a systematic analysis of a comparative and contextual type, thus allowing to overcome the limitations of classifications and analyses based on the subjective opinion of whoever reads and classifies the texts one by one. On one hand, the limits deriving from the expert reading of large quantities of texts is generally overlooked, even though they can produce significant distortions with effects in interpretation of the results. On the other hand, the adoption of automatic techniques of topic detection/clustering process of text analysis must be characterized by transparency in the specification of the methods of analysis and in the interpretation of the results, favouring their reproducibility both to qualify their scientific character and to favour their use in a systematic way over time or for corpora with similar characteristics. Along this direction, the paper suggests some key challenges to be made explicit in adopting topic models.

Data and methods

1.1 Data

The data in analysis are composed by a collection of news published in English by electrive.com, a daily newsletter covering a wide range of relevant information on developments in electric transport in Europe, the USA and China. As an exploratory step, we analysed the data source "electrive.com", provided as a service offered online by a private publishing company (Rabbit Publishing GmbH). It covers a wide range of relevant information on the developments in electric transport, and its daily newsletter is not only made available on the website, but is also relayed on the main social media, including Twitter.

Using the Twitter API, tweets from September 12, 2018 to March 8, 2022 were downloaded from the timeline of the electrive.com Twitter page. Within each tweet, we identified the link to the news URL. From the news page, with a web data extraction procedure (web scraping), we extracted the following information of each item of news: title, full text, associated tags, category, date of publication and links to the information sources.

Of the ten categories proposed by electrive.com - Air, Automobile, Battery & Fuel Cell, Energy & Infrastructure, Fleets, Politics, Short Circuit, Two-Wheeler, Utility

Vehicles, Water - the major category, "Automobile", encompasses nearly 38% of the news, followed by "Battery and fuel cells" and "Energy & infrastructure", each with nearly 14% of the news items.

1.2 Methods

The first step to be able to proceed to the analysis of the texts consists in structuring the textual information in a lexical and textual database. This step was carried out using TaLTaC2 software.

The electrive.com corpus is composed of 5,216 news items (title and full text) published in the period 12/09/2018-08/03/2022 and consists of a vocabulary of 54,230 different words (i.e. types) for a total size of 2,175,691 word occurrences (i.e. tokens).

By means of grammatical tagging of the vocabulary words, it was possible to distinguish between the different grammatical types of words (structure words versus content words) and also to lemmatize them, i.e. to relate each word to its canonical form, resulting in a reduction of the forms under analysis. Furthermore, thanks to the use of a lexical-textual model [14], it was possible to recognize the multiword expressions present in the texts. The recognition of these forms yielded lexical analysis units with less semantic ambiguity.

Thanks to the specific characteristics of news writing, it was also possible to distinguish easily between common nouns and proper nouns. In fact, the news was clearly and carefully written; use of uppercase and lowercase allows to identify proper nouns (of people and companies) and acronyms (defined by all capital letters). It was also possible to recognize all the words identified by the electrive.com magazine as TAGs of individual news items. At the same time, all the types (simple and compound) referring to nations (and national adjectives) mentioned in the text were identified.

In order to classify the news items on the basis of their similarity in terms of content, only common nouns (simple words and multiword expressions) and adjectives were selected for each news item. A vector space model representation was then generated, in which each news item is defined as a vector composed of the selected keywords. In the next step the matrix <news × keywords> (5,125 × 8,489) has been analysed through the different methods selected in order to define the topics covered within the corpus.

In addition to expert classification, three methods for clustering/topic detection have been implemented: LDA, Reinert method (ALC), Correspondence analysis and Cluster analysis (CA) to select the most relevant factors explaining variability within the corpus, on which a hierarchical cluster analysis is applied[1].

To compare the results across methods, we refer to two criteria. First of all, we check for semantic coherence in topic models [12, 17, 19], by applying two measures of coherence: the UMass score, based on a log-conditional-probability measure, and

[1] The following libraries have been implemented in R: *topicmodels* (for LDA), *FactoMiner* (for correspondence analysis), *quanteda* and *rainette* (respectively for, preparing the dataset and elaborating the Reinert method).

a variant of the UCI metric, based on the normalized pointwise mutual information. Both are intrinsic measures based on the co-occurrences in the corpus of the 10 most important words defined in each topic[2]. Secondly, we elaborate a cluster heat-map, to compare the results obtained by the different methods and the degree of association between the topics. Cramer's V was also used to measure the strength of association between the classifications produced by the different methods. With both criteria we refer to a given number of topics that is defined with the CA method.

Results: discussion and further developments

The first result refers to the optimal number of topics obtained with each method, in comparison with the 10 categories defined by the expert classification, with the three largest groups – "automobile", "battery and fuel cells", "energy & infrastructure" – encompassing, respectively, 35.5% and 14.1%, 13,8% of all news items.

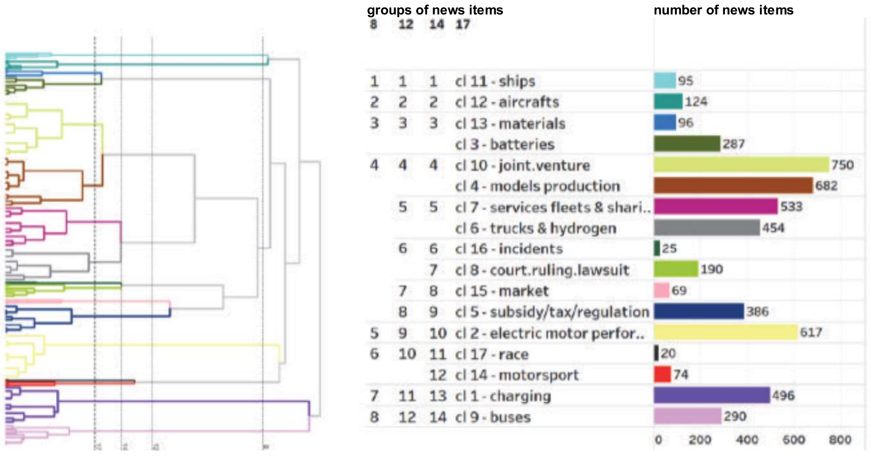
When considering the CA method, Figure 1 shows the dendrogram of the hierarchical clustering on the 10 factors of the correspondence analysis. The several cuts shown in the figure highlight the results from optimal number of clusters according to several methods (detailed results upon request). Our interpretation of results from an economic point of view indicates a cut at 17 clusters, which allows for a better disaggregation of the vast category "automobile" (split in the two groups of production differentiated with regard to features of economic organisation of production and a specific group describing electric motor performance), of the "battery &fuel cells" category (split in its components, respectively, of material and production), and of the" energy & infrastructure" (split in charging infrastructures vs. services). This number of 17 topics becomes the benchmark for all the other methods (details on optimal numbers are available upon request).

From the interpretative point of view of the topics encompassed in each cluster, the two methods that offer the hierarchical structure of texts of news items (CA and ALC) seem to be advantageous, since they allow us to define a greater or lesser number of topics, thus passing from a greater to a lesser detail, preserving the hierarchy of topics.

When moving on a more analytical comparison across methods of topic detection and documents classification, two key challenges must be addressed.

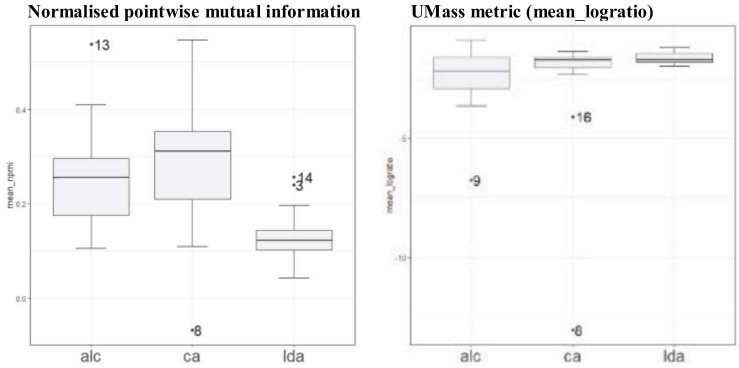
[2] The analysis in R uses the text2vec library. The selection of the terms identifying the topics is specific to the different methods used, respectively: test-value for CA; a chi-square test for ALC and the terms with the highest probability in the LDA.

Figure 1: Dendrogram: results of the CA method, with 8, 12, 14 and 17 groups



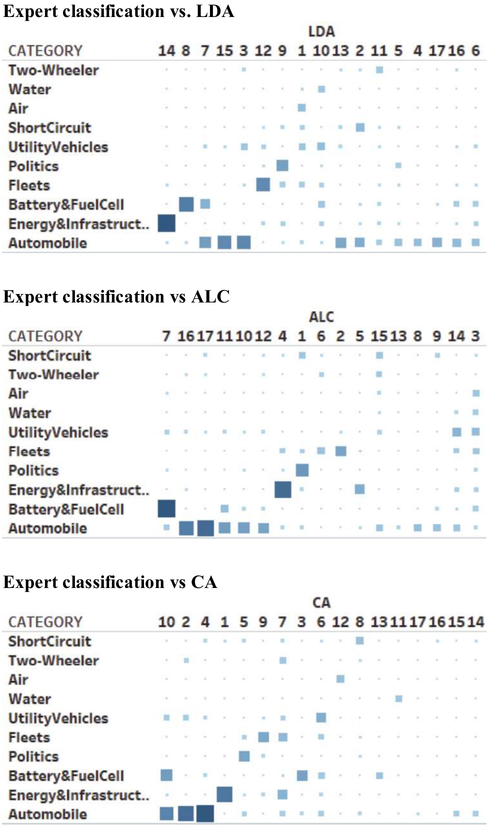
The first challenge concerns the ability of each algorithm to express semantically coherent topics. In this perspective, we implement two coherence indexes, both refer to a given number of topics that is defined with the CA method. Box plots in Figure 2 show the results of measures based, respectively, on normalized pointwise mutual information, NMPI, (left pane) and UMass score (right pane). According to NMPI, ranking of relative coherence shows the highest median value for CA method, with a high dispersion and cluster 8 as outlier that indeed has a miscellaneous of issues ("court.ruling.lawsuit"); with UMass score, LDA performs better, both in terms of median and overall topics, while in the case of CA method it highlights not only the case of cluster 8, but also of cluster 16, close by in the cluster.

Figure 2 – Box plots of coherence indexes



The other challenge refers to the metrics that can be used in comparing the results obtained with topic models under analysis, in terms of document classification. In this paper we use heat-maps (Figure 3) to visualise the results obtained by pairs of methods in groupings news items (in the 10 expert categories and 17 groups). We can observe that for some methods there are more areas of classification that overlap while in others overlapping is less significant.

Figure 2 – Expert classification (10 categories) and the three topic models under analysis (17 groups): heat-maps of relative correspondences



By comparing pairs of methods (results in Table 1) we observe a significant association between them (all p.values of the chi-squared test are <0.001). A summary measure of the strength of the association is provided by Cramer's V. It shows that CA is the methods that most closely approximate expert classification (67-68%), and is a confirmation of the superiority of the CA method in terms of readability of topics.

Table 1: Cramer's V indices for the three topic models

model.1	model.2	chi-squared	df	p.value	Cramer.V
CA	CATEGORY	21486.7	144	<0.001	0.6767
LDA	CATEGORY	14096.4	144	<0.001	0.5471
CA	ALC	24834.6	256	<0.001	0.5450
CA	LDA	24750.3	256	<0.001	0.5441
ALC	CATEGORY	12859.6	144	<0.001	0.5223
ALC	LDA	22410.6	256	<0.001	0.5174

Following Lebart [11] we intend to compare the results with other models (such as LSA and NMF), and to explore other models and methods for visualizing the comparative perspective on topic models and, in particular, Additive Trees, Self-Organizing Maps and Correspondence Analysis on the results of topic detection and clustering methods will be implemented. A third aspect to be explored is the analysis of the semantic similarity of the topics produced by the various algorithms. A fourt aspect concerns a general issue to be discussed, i.e. the specificity of cross-method results with respect to the characteristics of the corpus. In our database each document essentially deals with one topic and it would be important to discuss the comparison of topic models in cases of corpora with different structural features, in particular with regard to the variety of topics they might be include in each document.

References

1. Aggarwal, C.C., Zhai, C.: Mining text data. Springer Science & Business Media (2012).
2. Allan, J.: Topic detection and tracking: event-based information organization. Springer Science & Business Media (2012).
3. Berry, M.W.: Survey of text mining. Computing Reviews. 45, 9, 548 (2004).
4. Berry, M.W., Kogan, J.: Text mining: applications and theory. John Wiley & Sons (2010).
5. Blei, D.M. et al.: Latent dirichlet allocation. Journal of machine Learning research. 3, Jan, 993–1022 (2003).
6. Deerwester, S. et al.: Indexing by latent semantic analysis. Journal of the American society for information science. 41, 6, 391–407 (1990).
7. Feldman, R., Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge ; New York (2007).
8. Griffiths, T.L. et al.: Topics in semantic representation. Psychological Review. 114, 2, 211–244 (2007). https://doi.org/10.1037/0033-295X.114.2.211.
9. Hassani, A. et al.: Text mining using nonnegative matrix factorization and latent semantic analysis. Neural Computing and Applications. 1–22 (2021).
10. Lebart, L. et al.: Exploring textual data. Springer, Dordrecht; London (1998).
11. Lebart, L.: Looking for topics: a brief review. In: Text Analytics, Advances and Challenges. pp. 215–223 Springer (2020).
12. Mimno, D. et al.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 conference on empirical methods in natural language processing. pp. 262–272 (2011).
13. Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In: Text Analytics. pp. 17–28 Springer (2020).
14. Pavone, P.: Automatic Multiword Identification in a Specialist Corpus. In: Tuzzi, A. (ed.) Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. pp. 151–166 Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-97064-6_8.
15. Ratinaud, P., Marchand, P.: Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IRaMuTeQ. Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. 835–844 (2012).
16. Reinert, M.: "Alceste" - une méthodologie d'analyse des données textuelles et une application: "Aurelia" de Gerard De Nerval. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique. 26, 1, 24–54 (1990). https://doi.org/10.1177/075910639002600103.
17. Röder, M. et al.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp. 399–408 (2015).
18. Russo, M. et al.: Agents and artefacts in the emerging electric vehicle space. Int. J. Automotive Technology and Management. (2021).
19. Stevens, K. et al.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 952–961 (2012).
20. Sullivan, D.: Document warehousing and text mining. Wiley, New York (2001).