Neural Topic Modeling Project

Beatrice Alessandra Motetti
Politecnico di Torino
Student id: s287618
s287618@studenti.polito.it

Sara Moreno
Politecnico di Torino
Student id: s283832
s283832@studenti.polito.it

Stefano Galvagno
Politecnico di Torino
Student id: s290191
s290191@studenti.polito.it

Abstract

Topic models are a powerful tool to extract the main themes discussed in textual data. They can be improved by integrating, through knowledge distillation, the knowledge provided by a general language transformer-based model. In this paper, both an English and an Italian implementation are described, with the models trained on news data sets in the two languages. In addition, we evaluate qualitatively the topic model on out of domain articles. Furthermore we present a search engine system which, given a topic distribution of an article, provides the most related articles appearing in the training corpus.

1. Introduction

Neural topic models allow to extract from large amount of textual data the main themes discussed. Such models though suffer of lack of general knowledge, expressed in terms of a reduced representation of the documents over a bag-of-words. This means that related terms to a given words are not taken into account to extract the main topics of the document, while they could provide an extremely useful contribution in the task. To address this challenge, Hoyle et al. in [1], present a solution based on knowledge distillation. The main idea is that we could increase the knowledge of a neural topic model by integrating its document representation with some general language additional information provided by transformer-based language models. Such models are already pretrained, so they require only a fine-tuning phase before their integration with the base topic model. In this way, the general language knowledge of a huge model can be exploited to enrich the document representation of the topic model, by adding to its bag-of-words representation another distribution, which takes into account also non observed but related terms to the words. This concept is illustrated in Figure 1.

With these premises in this report we are going to describe the NTM replica based on knowledge distillation like in [1]. Furthermore we implemented with the same technique an Italian version and tested its performance with a qualitative analysis on out of domain articles. As last contribution we developed a pipeline which consists is passing a new article to the topic model (enriched with knowledge distillation) to extract its distribution over the previously learned topics and use it as input for a search engine system. As output it provides some documents of the training corpus which are mostly related to the new article provided.

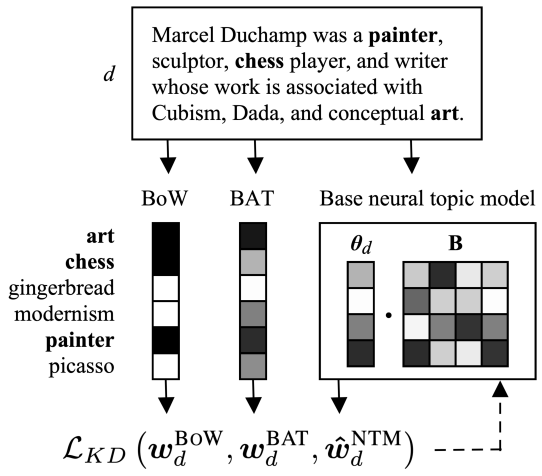

Figure 1: Image borrowed from [1]. It shows how the bag-of-words representation coming from the student model and the richer distribution provided by the teacher model are combined together in the loss function of the neural topic model.

2. Methodology

2.1. BERT

BERT (Devlin et al. 2017 [2]) is a transformer model capable of capturing deeper relationships between words thanks to its ability to consider both the left and right context of a term while producing its representation. BERT is pre-trained on huge text data sets in an unsupervised fashion with both masked language modeling and next-sentence

objectives. As output it gives a probability distribution over words taking into account non observed but related terms, outperforming the bag-of-word representation.

2.2. Topic models

LDA The LDA model is one of the oldest models used in natural language processing, proposed in the machine learning framework by D. Bley et al. in 2003 [3]. LDA has two inputs, α and β, that parameterize two Dirichlet distributions, from which we can derive two multinomial distributions, θ and φ respectively. By combining the outputs of the latter it is possible to obtain a set of words, those words are the only variables of the model that are not latent. During training the model iterates over all the words of all the documents multiple times, performing Gibbs Sampling.

SCHOLAR SCHOLAR is a topic model developed by Card et al. in 2018 [4] and it is based on a VAE. This model provides the possibility to incorporate both labels and covariates, but in our setting we disregarded it. It is important to highlight that the model treats each document as a bag-of-words representation.

The reconstruction error, which is computed considering both the bag-of-word representation (w_d^{BoW}) and the document reconstructed by SCHOLAR ($f(\theta_d, B)$), is the following:

$$\mathcal{L}_R = (w_d^{BoW})^T log\, f(\theta_d, B) \tag{1}$$

We decided to use the NPMI value as optimization metric because from a qualitative assessment we noticed that topics are more coherent with respect to the ones produced with perplexity as optimization metric.

2.3. Knowledge distillation

Knowledge distillation is a technique that allows transferring learning from a bigger model, called teacher, to a smaller one, called student. BERT is a reasonable choice as teacher thanks to its deep knowledge about language. It can thus enrich the student model by providing a more coherent and general document representation.

BERT comes pre-trained and the fine-tuning phase is left to the user since it depends on the specific downstream task it has to accomplish, in our case document-reconstruction. The fine-tuning process is carried out on the same data set used to train the student model.

BERT provides logits to the student, which is able to use them as soft labels and incorporate the latter in the loss function. Since we use SCHOLAR as student the loss function (1) is updated as follows:

$$\mathcal{L}_{KD} = \lambda T^2 (w_d^{BAT})^T log\, \hat{w} + (1-\lambda)\mathcal{L}_R \tag{2}$$

where λ and T are hyperparameters, w_d^{BAT} is the representation of document d provided by BERT and $\hat{w}$ is the original document representation provided by SCHOLAR scaled by the T factor. For more insights about the math refer to [1].

This approach is modular, since only few extra lines of code and three parameters are added to the original topic model.

2.4. Application: search engine

As a further extension, we developed a search engine that retrieves the documents in the training data set that are the most similar to those provided as input to the model. The only information needed by the search engine is the topic distribution of the article. It is possible to choose between two measures of similarity:

- *most probable topic*: the function retrieves the documents that have the highest value for the most probable topic of the article;
- *Jensen-Shannon divergence*: the selected documents are those that are defined by a similar topic distribution, according to the JS divergence, to the one that characterizes the article.

With the *Jensen-Shannon* option we can compare the entire distributions which is more effective when we have a document that is a mixture of many equally probable topics. It is due to remark the possibility that with the *topic* option, in case of an article with a predominant topic, the result can be mislead by a tail of small but not null probabilities.

3. Experimental setup

3.1. Baselines and model

As baselines we used two models, LDA and SCHOLAR, both presented in section 2.2. We then used a teacher model, BERT, that produces a document representation, to enrich the knowledge of the student model, SCHOLAR. As optimization metric for SCHOLAR we used the NPMI value.

3.2. Data sets

For the English replica we used the `20NG` data set. It is a collection of 18k documents divided into 20 classes, but since we are in an unsupervised setting we can leave this information out. For the Italian version we adopted the `webhose` data set [1]. This is a collection of news documents taken from the web, only a subset of about 15k items is used.

Preprocessing For the `20NG` data set we used the same preprocessing steps as the authors of the paper. They consist of fetching the data set from scikit-learn, both for the training and the test set. Then each document is tokenized, the

[1] `webhose.io/free-datasets/italian-news-articles/`

stopwords are removed and a lemmatization is performed. We take advantage of a clean version of the dictionary provided by Srivastava et al. (2017) [5]. Using that version, the vocabulary comprises 1995 words.

For the LDA model [3], we applied the same preprocessing steps, tokenizing, removing stopwords and lemmatizing each document. We also performed an additional mapping step to match the LDA input format which is different form the SCHOLAR's one.

Given the Italian data set, we tokenized each document, removed the standard Italian stopwords extended with some recurring words that would affect the topics if not removed. Finally, we lemmatized each document using the `treetagger` tool that provides an Italian lemmatizer. After that we perform a vocabulary reduction adding the constraint that a word can be present in the dictionary only if it appears in at least 20 documents and in less than 65% of them. In that way the vocabulary is composed of 10083 words.

3.3. Evaluation metrics

Internal and external NPMI To evaluate the quality of the learned topics we decided to use the Normalized Pointwise Mutual Information (as in [1]). The NPMI metric is useful to measure the probability of co-occurrence of two words with respect to their single occurrences in the corpus. Thus it is effective to quantify the coherence of the words which mostly define a topic and evaluate its quality. It is possible to compute both an internal NPMI and an external NPMI value, as in [4]. The internal NPMI takes the co-occurrences counts from a subset of the data set on which the training is based, thus it is more specific to the task. The external NPMI is instead computed on top of a different corpus.

Palmetto evaluation framework To obtain the external NPMI of the topics, the Palmetto evaluation framework has been adopted [6]. With this tool, it is possible to choose between different coherence measures; to remain consistent with [1], the NPMI has been used. The corpus on which Palmetto computes the occurrences and co-occurrences of the topic pairs of words is the English Wikipedia. This approach is modular: we can obtain the evaluation results for all the topic models by using only their list of topics, independently of the model.

Topic alignment procedure It is also possible to perform a topic alignment technique, in order to be able to compare the topics extracted by SCHOLAR and SCHOLAR+BAT directly. We followed the same procedure described in [1], where the pairs of topics (one of SCHOLAR and one of SCHOLAR+BAT) are aligned based on their similarity in terms of words distributions. To compute this distance, the Jensen-Shannon divergence is used.

K	NPMI	LDA	SCHOLAR	SCHOLAR+BAT
50	Internal	0.1759	0.3270	**0.3402**
	External	-0.0041	0.0318	**0.0617**
200	Internal	0.1679	0.2617	**0.2792**
	External	-0.0530	0.0075	**0.0557**

Table 1: Internal NPMI values obtained with the different models and number of topics on the `20NG` data set.

4. Results

4.1. English replica

The results show a consistent improvement of the NPMI value with the usage of knowledge distillation with respect to the LDA and SCHOLAR models on the `20NG` data set, as reported in Table 1. Considering both 50 and 200 topics, we can see that SCHOLAR+BAT outperforms the other baselines in terms of both internal and external NPMI. It is interesting to observe also that a smaller number of topics yields a higher NPMI value. To better understand this result, we performed additional runs of the SCHOLAR+BAT model with $K = \{10, 25, 75\}$. We could not test more values due to computational constraints. The obtained results, reported in Figure 3, show a decreasing trend of the NPMI value when the number of topics increases.

It is possible to have an overlook on the topics distribution by using the visualization tool `pyLDAvis` [7]. It provides an overview of the topics and their differences by plotting them as in Figure 2. The area of a circle, i.e. a topic, represents its relative frequency in the corpus. The closer two topics are, the more similar they are. An interesting feature provided by this tool is to highlight the conditional distribution of a word over topics, in order to understand if such term tends to appear in more than one topic. In Figure 2 for example, the word *drive* has been selected and it is possible to see that it is relevant in mainly two clusters of topics, one at the bottom right and one on the middle of the graph. Upon a manual inspection of the most relevant words of the involved topics, we can understand that one cluster of topic is related to the computer and electronics field (for example topic 39 in Figure 2), while the other one refers to cars and motorbikes. Thus, it is intuitively correct that the word *drive*, due to its polysemy, appears in these topics related to completely separated semantic fields.

We can have a look at the trend of the internal NPMI value in each epoch during the training process of the student model for each chosen number of topics (Figure 4). From these results we can see that after a major peak the NPMI value tends to decrease because of model's overspe-

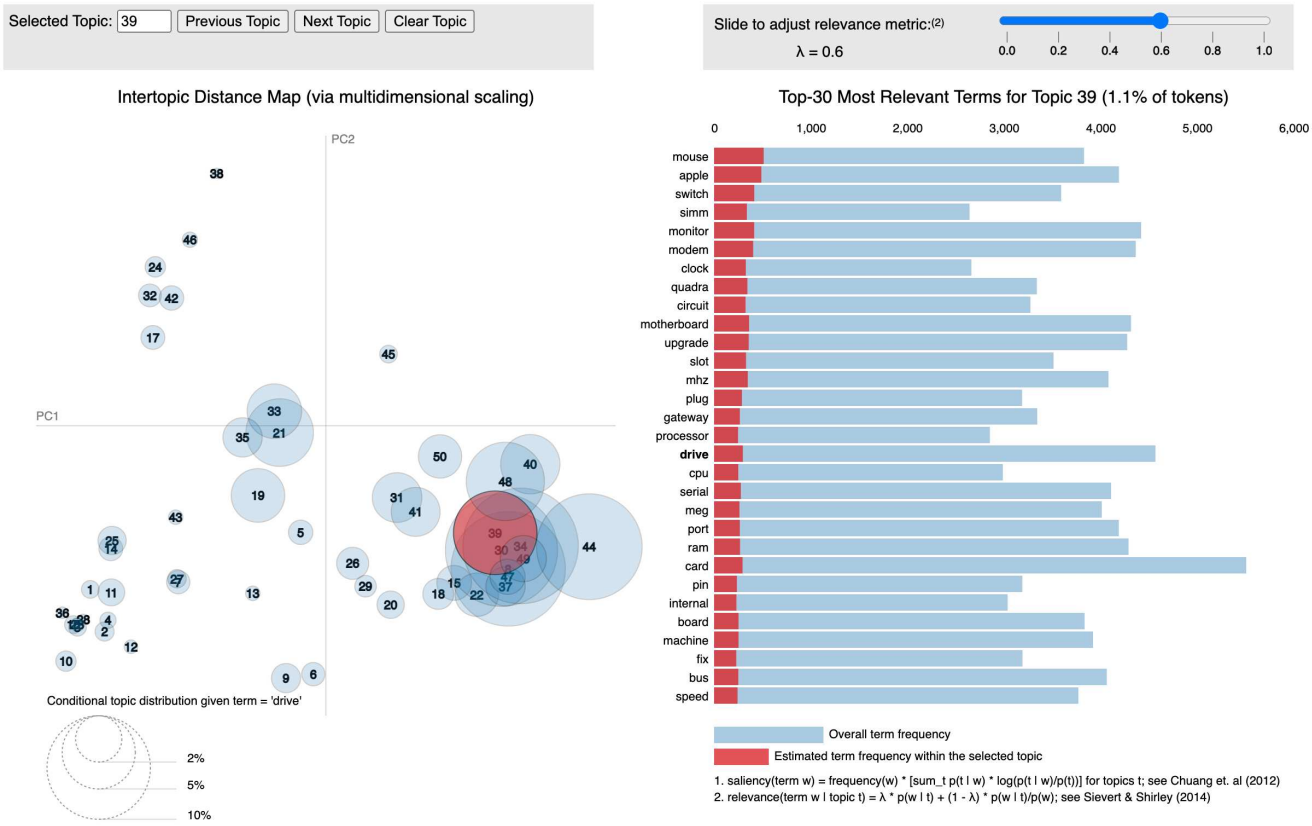

Figure 2: Map of the extracted topics, most relevant words for a randomly selected topic (on the right) and conditional distribution over topics of the word *drive*.

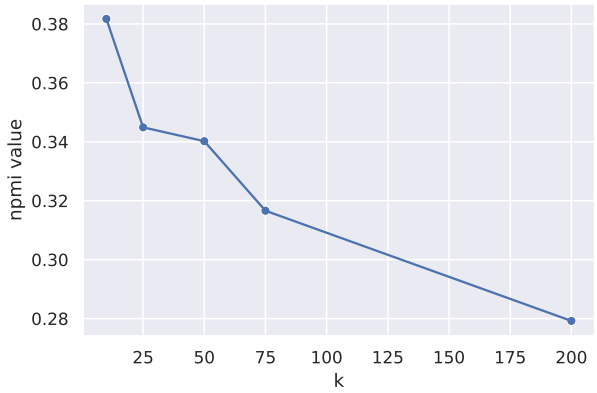

Figure 3: NPMI values for each tested number of topics.

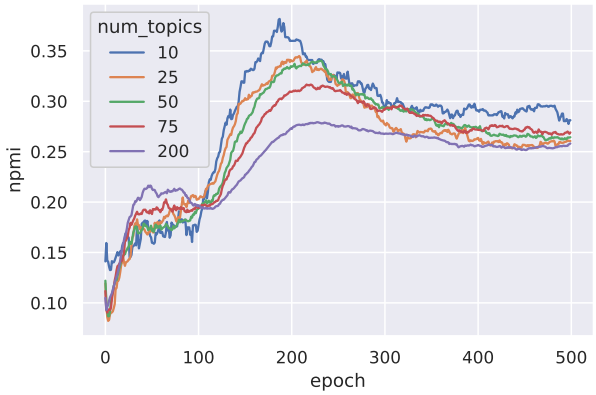

Figure 4: Trend of the NPMI value over the training epochs of the student model for all the different number of topics which have been tested.

cialization.

We can thus perform the topic alignment technique described in 3.3. In Table 2 some examples of aligned topic pairs are shown.

4.2. Italian version

As first step we performed several trials, each with a different learning rate, and chose the one that per-

JS divergence	Model	Topic	Internal NPMI
0.0200	SCHOLAR	program file byte contest postscript character input info remark exit	0.1123
	SCHOLAR+BAT	file echo contest remark title int exit larry input section	0.1149
0.0244	SCHOLAR	teaching islamic church islam religion catholic muslims muslim tradition marriage	0.3488
	SCHOLAR+BAT	teaching church catholic marriage tradition doctrine faith jesus christianity religion	0.4426
0.0320	SCHOLAR	orbit rocket spacecraft space fuel solar flight mission moon vehicle	0.4008
	SCHOLAR+BAT	spacecraft space orbit mission rocket shuttle solar nasa satellite lunar	0.4378

Table 2: Examples of aligned topic pairs.

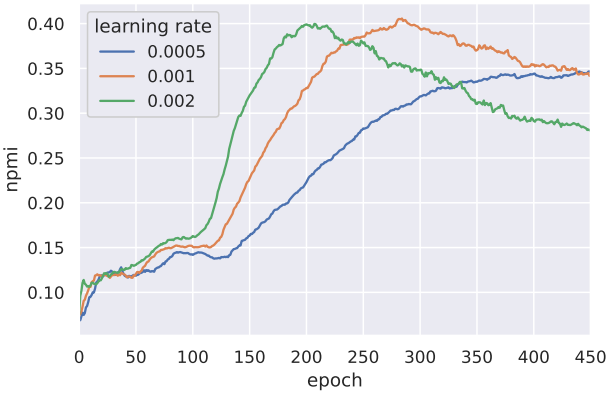

Figure 5: NPMI trends with different learning rates.

K	NPMI	LDA	SCHOLAR	SCHOLAR+BAT
50	Internal	0.2138	0.4025	**0.4057**

Table 3: Internal NPMI values obtained with the different models and number of topics on the Italian data set. Note that the external NPMI is not computed because the Palmetto evaluation framework is available only for the English language.

forming better in terms of NPMI. We tested as values $\{0.002, 0.001, 0.0005\}$ and we got the highest NPMI value (0.4057) with a learning rate of 0.001 (Figure 5).

The results obtained with the models trained on a data set in Italian are consistent with those obtained on the 20NG and are reported in Table 3. LDA performs substantially worse than SCHOLAR, which also yields a NPMI value lower than the one obtained by using knowledge distillation. Nevertheless, the gap between the NPMI value obtained by using SCHOLAR and SCHOLAR+BAT is lower than the gap obtained in English on the 20NG data set.

We can perform the topic alignment with the same technique used for the English replica. In Table 4 some examples of aligned topic pairs are shown.

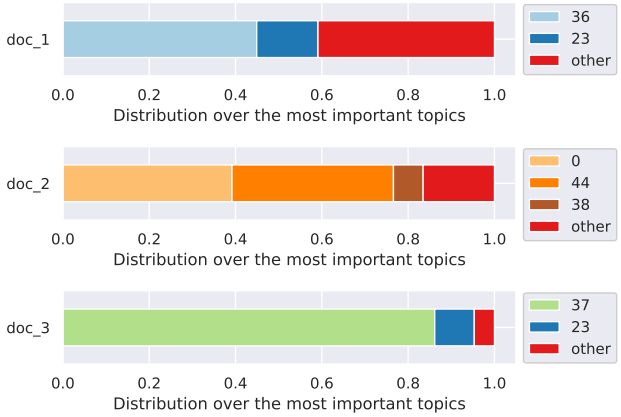

Figure 6: Distribution over the most important topics for each document. The topics having a probability less than *0.05* have been collapsed into a general topic called "other". See Appendix B for more precise results.

4.3. Qualitative analysis

To evaluate qualitatively the model, we can provide some new articles to the topic model and extract their topic distributions. To obtain the document representation, both the teacher and the student model are employed without performing any further training of course, in order not to modify the learned knowledge. This can be done both with the English and the Italian replica, but since our main contribution is the Italian version, we decided to present only those results on behalf of conciseness. We randomly selected three articles published by *Corriere*, an important Italian newspaper. The first concerns some Italian start-ups in the health sector, the second the Afghans' exodus after the Taliban capture of Kabul in August 2021 and the third the importance of vaccines to limit the spread of Covid-19.

The results of the qualitative analysis are shown in Figure 6. It is immediate to notice that all the articles are characterized by a dominant topic, having the majority of the probability mass itself. We can also see how each document is mostly defined by two or three relevant topics; the others have been assigned a negligible probability by the topic model. In Figure 6, such irrelevant topics have been

JS divergence	Model	Topic	Internal NPMI
0.0324	SCHOLAR	politica democrazia politico confliggere libertà toga lobby altrove culturale sociale	0.1590
	SCHOLAR+BAT	democrazia politica popolo politico religioso dignità principio comunità confliggere libertà	0.2519
0.0339	SCHOLAR	wall ftse fca mib analista rialzo street listino decennale bund	0.4243
	SCHOLAR+BAT	analista fca wall trimestre trimestrale ftse rialzo ribasso dollaro listino	0.4266
0.0671	SCHOLAR	malattia paziente patologia farmaco terapia pazientare vaccino medicina vaccinare ricercatore	0.4206
	SCHOLAR+BAT	paziente vaccino vaccinare malattia vaccinazione patologia farmaco trattamento salute medico	0.4540

Table 4: Examples of aligned topic pairs.

collapsed into a general "other" category.

It is interesting to notice how the model in capable of generalizing since it retrieves a meaningful and coherent topic distribution even for articles characterized by very recent and unforeseen themes. For instance with the article about Covid-19 we obtain as most representative topic with probability 0.86 the following one: [*paziente, vaccino, vaccinare, malattia, vaccinazione, patologia, farmaco, trattamento, salute, medico*].

4.4. Search engine

As in 4.3, we will present only the results obtained with the Italian replica. Note that the technique is identical for the English version. These results have been evaluated only from a qualitative perspective. We tested the search engine with the same documents used in 4.3. The outputted articles are consistent and coherent in terms of topics with the input ones. The *topic* option performs better than the *Jensen-Shannon divergence* option, this is probably due to those non-relevant topics which do not individually define the document, but all together assume a consistent percentage of probability mass and thus play a significant role in the JS computation.

5. Related works

Hoyle et al. presented in 2020 a paper [1] showing how knowledge distillation could bring a great improvement on topic models' performances and that's the work that inspired our project.

As previously mentioned in 2.1, BERT (Devlin et al., 2017 [2]) has played a very important role for the work presented in this report.

For what concerns the topic models we referred to SCHOLAR [4] and LDA [3].

6. Conclusions

The obtained results show how knowledge distillation is a working strategy which is able to improve the performance of a topic model. This happens both for the English and the Italian version, even though we noticed that in Italian the gap in the NPMI value between SCHOLAR and SCHOLAR+BAT is less than in English.

As future works, it would be interesting to see the impact on the results of the adoption of an additional preprocessing step to compute the tf-idf and select the words to include into the vocabulary according to that value as well. This might increase the quality of the vocabulary and thus, indirectly, also the one of the extracted topics. Furthermore, a bigger training data set (which we could not use due to computational and memory constraints), might as well improve the knowledge gained by both the teacher and student model. It would also help to have more recent and various documents, in order to have better search engine results.

References

[1] Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. Improving Neural Topic Models using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1752–1771, Online, November 2020. Association for Computational Linguistics.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] Dallas Card, Chenhao Tan, and Noah A. Smith. Neural models for documents with metadata. In *Proceedings of ACL*, 2018.

[5] A. Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *ICLR*, 2017.

[6] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6*, 2015.

[7] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA*, pages 63–70. Association for Computational Linguistics, 06 2014.

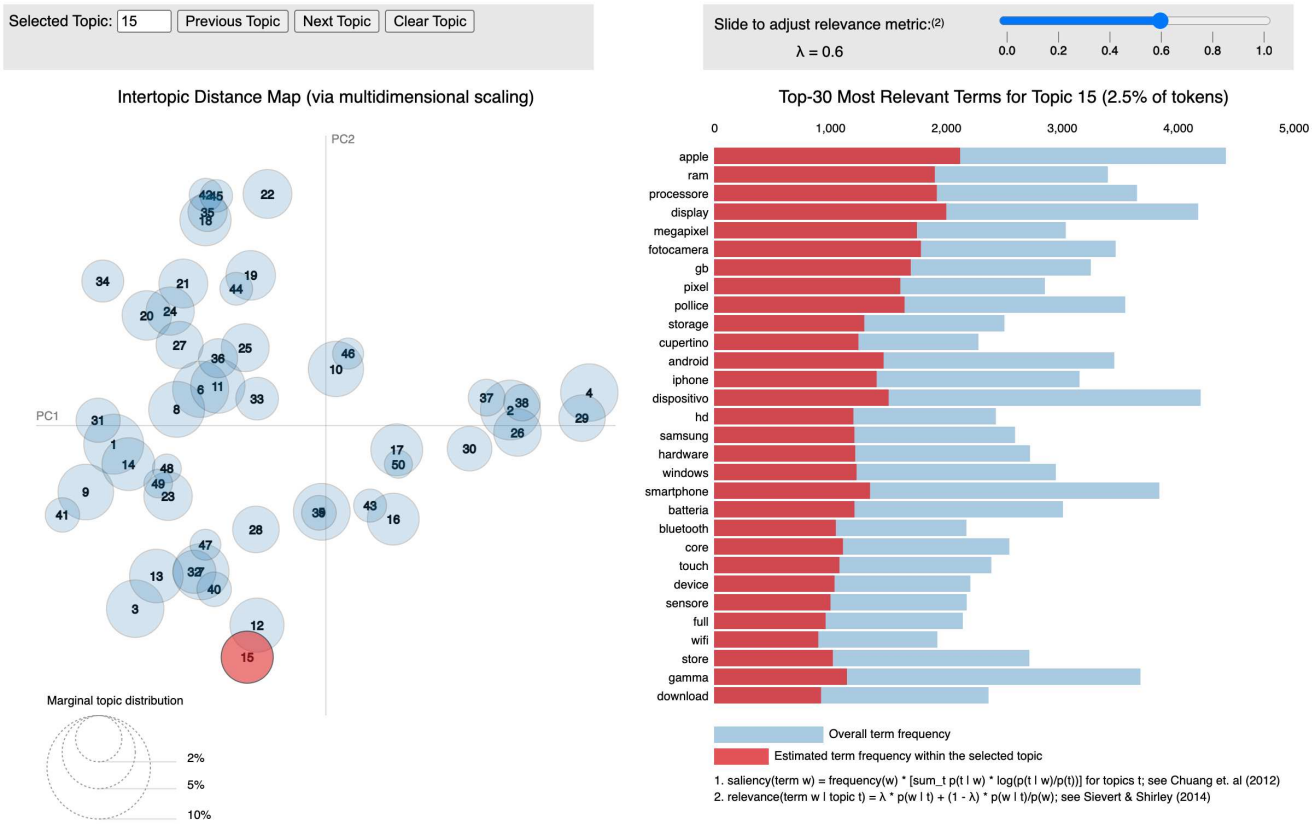

Figure 7: Map of the extracted topics, most relevant words for a randomly selected topic (on the right).

Topic id	First 10 words
0	'siriano', 'siria', 'assad', 'bashar', 'libia', 'mosca', 'turchia', 'putin', 'russia'
23	'ricercatore', 'dieta', 'alimento', 'carne', 'cancro', 'scienziato', 'proteina', 'verdura', 'sviluppare'
36	'innovazione', 'innovativo', 'business', 'infrastruttura', 'sostenibilità', 'tecnologia', 'strategico', 'industria', 'tecnologico'
37	'paziente', 'vaccino', 'vaccinare', 'malattia', 'vaccinazione', 'patologia', 'farmaco', 'trattamento', 'salute'
38	'palestinese', 'israeliano', 'siriano', 'islamico', 'israele', 'abu', 'siria', 'aereo', 'cisgiordania'
44	'democrazia', 'politica', 'popolo', 'politico', 'religioso', 'dignità', 'principio', 'comunità', 'confliggere'

Table 5: Main topics present in the articles from *Corriere*.

Appendix

A. `pyLDAvis` visualizations

We include in the appendix the visualization of the topics made with the tool `pyLDAvis` on the Italian model with knowledge distillation and $K = 50$ topics. In Figure 7 a random topic has been selected, in order to show its main words.

B. Qualitative analysis on the articles

We reported in Table 5 the first 10 most important words for each of the topic mentioned in Figure 6 in order to allow a better understanding of our analysis.

C. Search engine results

We further reported in the appendix some results of the search engine. We focused our attention on the Italian version. In case of the *JS divergence* option, the value in each row of Table 6 corresponds to the distance between the topic distribution of the article and the one of the retrieved document. In case of the *topic* option, the value returned in every row of Table 7 is the proportion of the most probable topic of the article in the retrieved document. It is possible to notice how the documents resulting form the search engine in this second case (Table 7) are more similar to the articles in input, how we explained in 4.4.

Article's title	JS divergence	Retrieved document's title
Quattro start-up italiane premiate nella prima edizione dell'Health&BioTech Accelerator[2]	0.1024	Kiron, la StartUp etica per l'istruzione degli immigrati[3]
Afghanistan, il commissario Onu: "Chi resta avrà bisogno di aiuti. Esodo? Sarà regionale"[4]	0.1344	In Tunisia e Libia più dialogo meno forza [5]
Locatelli: "Covid, circa il 12% dei vaccinati può infettarsi, ma non sviluppa la malattia"[6]	0.0542	Il calcio coronarico dimezza il numero di candidati delle statine [7]

Table 6: Articles retrieved by the search engine using the *Jensen-Shannon divergence* option

Article's title	Topic prevalence	Retrieved document's title
Quattro start-up italiane premiate nella prima edizione dell'Health&BioTech Accelerator	0.9814	SAIE 2015. Successo della formula Smart House [8]
Afghanistan, il commissario Onu: "Chi resta avrà bisogno di aiuti. Esodo? Sarà regionale"	0.9668	Un mini-summit sulla rotta dei Balcani (flusso dei profughi) [9]
Locatelli: "Covid, circa il 12% dei vaccinati può infettarsi, ma non sviluppa la malattia"	0.9897	La Campagna Vaccinale Antinfluenzale [10]

Table 7: Articles retrieved by the search engine using the *Topic* option

[2] Quattro start-up italiane premiate nella prima edizione dell'Health&BioTech Accelerator

[3] Kiron, la StartUp etica per l'istruzione degli immigrati

[4] Afghanistan: "Chi resta avrà bisogno di aiuti. Esodo? Sarà regionale"

[5] In Tunisia e Libia più dialogo meno forza

[6] Locatelli: "Covid, circa il 12% dei vaccinati può infettarsi, ma non sviluppa la malattia"

[7] Il calcio coronarico dimezza il numero di candidati delle statine

[8] SAIE 2015. Successo della formula Smart House

[9] Un mini-summit sulla rotta dei Balcani

[10] La Campagna Vaccinale Antinfluenzale