Topic Modeling on Italian Dataset

Federico Germinario
Politecnico di Torino
s282869@studenti.polito.it

Filippo Buffa
Politecnico di Torino
s290460@studenti.polito.it

Gabriele Gioetto
Politecnico di Torino
s285501@studenti.polito.it

Abstract

Topic modelling is one of the main solution to link large collection of documents to generated topics.
Using an italian version of a pretrained transformer, we exploit knowledge distillation with a probabilist topic model.
We show an improvement in coherence both on average over all the topics and in aligned pairs with respect to baseline models.
We propose a visualization of the topics using t-SNE and we test our model on new unseen articles from an italian newspaper.

1. Introduction

Modeling the topics of a corpora of documents is one of the main tasks of NLP. In the past twenty years many models have been proposed. Some of them treated documents just as group of tokens, others as a probability distributions over topics, some were able to incorporate metadata such as ratings or the author. In this paper we will follow the procedure suggested by Hoyle *et al*. in 2020 [4], first we will reproduce their experiments on the 20 newsgroup dataset and than we will propose a model for the Italian language.

The proposed model exploits the language knowledge of a pretrained transformer (BERT) to guide a student model using knowledge distillation.
Using BERT as a teacher, we can exploit a dense description in the task of generating the prediction of a topic over the distribution, including terms that are not observed in the documents.

A model for the Italian language is proposed and compared with baseline models, using as a measure the topic coherence. A test on several articles of "La Stampa" is then presented, to test the goodness of the topic modelling on external documents, applying a dimensionality reduction over the topics distribution, in order to provide a visualization of the articles.

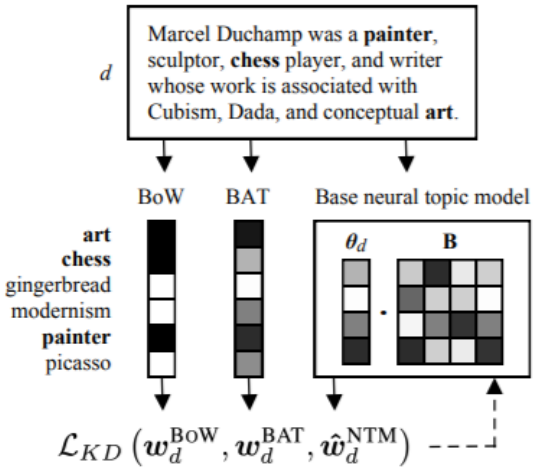

Figure 1. Example of how BAT provides a knowledge distillation. Image taken from Hoyle *et al*. [4]

To summarize our main contributions:

- We replicate the model on the 20 newsgroup dataset
- We implement an Italian version of BERT-based Autoencoder as Teacher
- We compare the proposed model, achieving better results with respect to other state-of-the-art models
- We propose a visualization and inference on articles from an external datasets, giving a visualization of the articles distribution over the topics.

2. Related works

Different topic models have been presented, they are based on the assumption that each document consists of a mix of topics, which can be described by a collection of words, we present some of the main related works regarding topic models **LDA**: since it was presented in 2003 by Blei [1], this method has been studied and implemented in

various forms due to its power and the interpretability of the output. The main idea behind it is to represent documents as a random mixture over latent topics and topics as a mixture over words, both are assumed to be distributed as a Dirichlet distribution. It has a generative part, in which a topic is sampled and than a word is sampled conditioned to the topic. This method does not have a loss function, the training by Bayesian methods.

SLDA: introduced by McAuliffe and Blei in 2008 [5], this is an improved version of LDA, the structure is the same, but a response variable has been added to each document. This parameter can be seen as a metadata, for example the date or the rating of the article. Also a variational loss function has been introduced to train the model.

SCHOLAR: this model is the basis of our work, it has been introduced in 2018 by Card *et al*. [2]. Its main role has been to introduce a better way to incorporate metadata into the analysis and they also proposed a *logistic normal* prior instead of the *Dirichlet*. It is possible to take this model into SLDA by changing the prior and ignoring covariates.

Transformers are widely used in various NLP tasks, they are models trained on large corpus capable of exploiting a rich knowledge of the language. In our case the task of knowledge distillation is assolved by two different versions of BERT.

BERT: maybe the most game changing model in NLP in the past years, it has been presented by Devlin *et al*. in 2018 [3]. It is a transformer-based model. It has been used to solve various kind of task. The main idea behind is to treat the text in a bidirectional way.

3. Methodology

In our approach, in order to generate the topics distributions in the training step, we adopt the SCHOLAR framework. We will give a briefly description of SCHOLAR in the next sections, but one of the main differences to point out between our adopted baseline LDA, consists in having a loss function:

$$\begin{aligned}\text{ELBO} = \mathbb{E}_{q(\theta_d|\cdot)}\left[\mathcal{L}_R\right] \\ - \mathbf{D_{KL}}\left[q\left(\theta|w^{BoW}, x_d\right)||p\left(\theta_d\right)\right].\end{aligned} \quad (1)$$

Where θ_d is represent topic distribution, $q(\theta_d|\cdot)$ is the variational distribution, $p(\theta_d)$ is the posterior distribution, w^{BoW} is the bag of word representation of the document, $\mathbf{D_{KL}}$ is the Kullback–Leibler divergence which gives a measure of similarities of the two distributions and

$$\mathcal{L}_R = (w_d^{BoW})^T log\left(f\left(\theta_d, B\right)\right) \quad (2)$$

with $f\left(\theta_d, B\right) = \sigma\left(m + \theta_d^T B\right)$. The function f is the multinomial distribution over words, σ is the softmax function, m is a prefixed background over words and B is a matrix where each row corresponds to the k-*th* topic-word probabilities in log-frequency space.

3.1. Knowledge distillation

Knowledge distillation, the key process adopted, is the procedure where two models are used together sequentially. The first one is trained and used to take some information that were not explicit in the initial dataset. This new knowledge is passed to the second model, which is the actual model that gives the desired output. This can be seen as a complicated preprocessing, in our case the reason why it is done is to be able to pass to the main model a more general picture.
We use **BAT** as teacher, which pass to the student (**SCHOLAR**) an $N \times V$ matrix, where N is the number of documents and V the vocabulary size. This matrix has the probability that each word appears into the text. We can reformulate $\mathcal{L}_\mathcal{R}$ that we find in (1) taking into account the information provided by the teacher

$$w^{BAT} = \sigma\left(z_d^{BAT}/T\right) N_d \quad (3)$$

$$\hat{w} = f(\theta_d, B; T) \quad (4)$$

$$\mathcal{L}_{\mathcal{KD}} = \lambda T^2 (w_d^{BAT})^T log\left(\hat{w}\right) + (1-\lambda)\mathcal{L}_R \quad (5)$$

where z_d^{BAT} are the logits produced by the teacher, T is the softmax temperature, which is a parameter that controls the estimated probabilities over the words. It is easy to see that the new $\mathcal{L}$ is a convex combination of the previous and a new term associated to the knowledge provided by the teacher.

3.2. BAT

BAT is the acronym for **B**ERT-based **A**utoencoder as our **T**eacher. As explained in the previous part this bigger model plays the role of guide for the student, giving it more information. In our cases it is capable to give information about the context. Indeed, instead of passing the pure bag of word representation of the document, thanks to BAT, we are able to give a weight to the words base on their coherence in the text. For example you can see in figure 1 that the teacher is able to understand the main topic and gives an higher score to words correlated to it even if they are not present in the text.
This teacher is based on a distilled version of BERT, which is a powerful multi-layer bidirectional Transformer encoder introduced by Devlin *et al*. in 2019 [3]. BERT stands for **B**idirectional **T**ransformers for **L**anguage **U**nderstanding. The training part of this model is divide into two subparts. First there is a pre-training and than a fine-tuning. In the former, first the text is divided into sentence with length less than 512 and than a masked language model is used. In this procedure 15% of the words are randomly masked and the model is trained to recognize them, the fact that it is

an undirectional procedure make the BERT bidirectional in this sense. The last part of the pre-training is next sentence recognition. A sentence is taken and another is proposed to the model which is trained to recognize if it is the following, the second sentence is taken with 50% probability to be the real next one. To fine-tune the model uses the self-attention mechanism to unify these two stages, as encoding a concatenated text pair with self-attention effectively includes bidirectional cross attention between two sentences. The version we are using is a distilled from of BERT introduced by Sanh *et al*. in 2019 [8]. It smaller and faster compare to the original one, which has been trained on very large batches leveraging gradient accumulation taken from the actual BERT. This type is smaller but retains 97% of the BERT's performances.

3.3. SCHOLAR

As we said we adopted SCHOLAR as our topic model combined with BAT. Presented by Card *et al*. 2018 [2], the main achievement of this tool is the incorporation of metadata into the analysis. The choice of this student has been made based on its capability on working on preprocessed data. Indeed with other type of tools like **LDA**, due to the absence of loss function, is more difficult to analyze other type of data. The generative story of the used **SCHOLAR** is the following.

For each document i of length N_i:

Draw a latent representation of the document into the simplex trough a logistic normal prior:

$$r_i \sim \mathcal{N}(r|\mu_0(\alpha), diag(\sigma_0^2(\alpha)))$$
$$\theta_i = softmax(r_i)$$

Generate words:

$$\eta_i = \theta_i^T B$$

For each word j in document i:

$$w_{ij} \sim p(w|softmax(\eta_i))$$

Generating labels:

$$y_i \sim p(y|f_y(\theta_i))$$

where B has a Dirichelet prior, w_{ij} and y_i are distributed as multinomials and f_y is a multi-layer neural network. The parameters $\mu_0(\alpha)$, $(\sigma_0^2(\alpha)$ are taken form a multivariate normal prior. As in others works like **LDA**, each document is assumed to have a latent distribution over topics (i.e. r_i). To be able to approximate a posterior over the topics the sampling-base VAE framework has been used. The used of this tools took to the loss function (1), which as explained before it has been modified based on information from the teacher. This model is capable of inferring latent topics (*i.e*. it deduces from the corpus a word representation of topics), infer a latent representation for new document and classify them trough

$$\hat{y}_i = \arg\max_{y \in Y} p(y|r_i). \tag{6}$$

3.4. Coherence

Coherence measures are methods that can be used to evaluate the goodness of a topic proposed by the model. The model is capable to give a representation of labels trough the set of most important words. Topics produced in this way are not guaranteed to have a good interpretability, to calculate this many tools come to help. In our case we decided to use NPMI which stands for Normalized Pointwise Mutual Information. First let us introduce PMI

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}. \tag{7}$$

This measure is capable to give a score the co-occurrence of two words. The probabilities $P(x)$ and $P(y)$ are the frequencies of the words x and y into a given corpus and $P(x,y)$ is the frequencies where both of them appears in the same document. To calculate the NPMI we have to introduce another term

$$h(x,y) = -\log_2 P(x,y), \tag{8}$$

which gives the log frequency of the co-occurrence. Dividing (7) by (8) we obtain NPMI

$$NPMI(x,y) = \frac{PMI(x,y)}{h(x,y)}, \tag{9}$$

which is bounded in $[-1,+1]$. A score of -1 stands for no co-occurrence and $+1$ means that the two words always appears together.
Palmetto is the framework we used to calculate explicitly our scores. It has been presented in 2015 by Röder Michael [7] . Its base corpus is Wikipedia. We also calculate the internal NPMI, which is simply the score calculate on the dataset we used.

3.5. Pseudo-code

In this part we want to highlight and summarize our procedure.

Algorithm 1 Procedure

procedure PREPROCESSING
    download dataset and vocabulary
    removing stopwords
    lemmatization
    tokenization
procedure TEACHER
    download vocabulary (.json)
    download word counts and id of documents
    training model (AdamW method)
    calculate logits of words
procedure STUDENT
    load vocabulary, id, word counts
    load logits from the teacher
    filter null documents
    keeping k words with highest logits scores
    for each word calculate the frequency
    load the word2vec vectors
    replace the randomly initialized vectors with word2vec
    initializing and training model
loop over batches:
    compute accuracy and average loss on minibatch
    update the priors on the individual weights

4. Experimental setup

We focused on two different models, one for the English language and a second one for Italian. Both the models follow the same approach with a different dataset and a different bert model implemented.

4.1. Dataset and metrics

For the english model 20 newsgroup has been proposed as dataset. It is a well known collection of documents used in text classification and clustering. It consist of nearly 18000 documents on 20 topics split in training and testing subsets. The implementation of BAT + SCHOLAR for the Italian language has been made training on the "Italian news articles dataset" provided by webhose.io. The dataset is composed of 159,226 articles, however we used a subset of 20200 articles, a number similar to the size of the 20 newsgroup dataset, to speed up the training process. After training our model, we have tested it on 20 articles of the "La Stampa" newspaper, not belonging to the original dataset. Ten of these articles belong to the "Sport" section, while the other ten belong to the "Economy" section.

In order to evaluate the coherence of the topics, NPMI is used as metric taking the top ten words for each topic and taking the average over all the topics. This metric is computed both on the internal corpus and then externally using Palmetto through its web application.

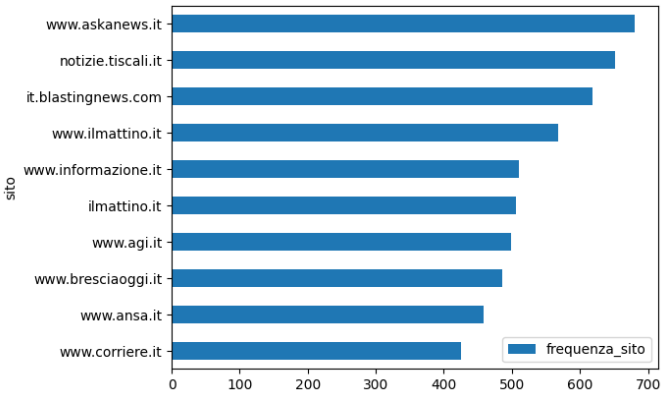

Figure 2. Most frequent sources of the articles in the webhose.io dataset

4.2. Preprocessing

For the preprocessing we have replicated the 20 newsgroup's one, given that the formats of the two datasets were similar. To create the vocabulary, we have adopted the same processing used to create the 20 newsgroup vocabulary. It includes:

- Removing punctuation and symbols
- Removing numeric values
- Removing Italian stop words
- Keeping words that appear at least on 10 documents and in less than the 70% of the documents

The scholar model remains the same, since it only needs the logits to work, independently from the language of the words and articles corresponding to the logits.
Once the model was trained and the topics were created (K = 50), we used the scholar model just obtained to associate each "La Stampa" article to 50 scores, in which the ith-score is the probability that the article is about the ith-topic.

4.3. DistillBERT models

The pretrained version of BERT used in the teacher for the english model is the bert-base-uncased one, in which capital letters are converted in lowercase, before the model is applied. For the italian model, instead of a multilingual model, we adopted a model pretrained on a large corpus of only italian documents, uncased[1].

[1]The dbmdz/bert-base-italian-xxl-uncased from huggingFace.co, trianed on a corpus of 13,138,379,147 tokens from OSCAR and OPUS corpora

Section	Most frequent Topic
Economy	equita dax analisti warning trimestrali
Sport	allegri tuttosport bonucci khedira champions

Table 1. First 5 words of the words of the most frequent topics for the Economy and Sport articles

5. Results and discussion

As we said two different models are generated, one for the English version trained on the 20 newsgroup corpus and the second one trained on the Italian corpus, previously described.
An internal and external NPMI are computed for all the topics for the English version, while for the Italian version, only the internal NPMI is provided.

NPMI is computed both as average over all the topics, and for the single topic. For each topic we operate an alignment pairing topics from SCHOLAR and SCHOLAR + BAT using the Jensen-Shannon divergence for both the Italian and English corpora.

Lastly, we propose a classification of the 20 Italian articles, computing the probability distribution of an article belonging to a topic, we provide the most frequent topic for each major category ("Sport" and "Economy") (Table 1) and we show the plots of the word embeddings. These are obtained using PCA dimensionality reduction and the fastText Italian word embeddings.

5.1. Impact on group coherence

Comparing BAT with LDA and SCHOLAR, we can appreciate an overall better coherence in terms of group topics. Table 2 shows that BAT outperforms both the baselines and achieves overall better coherence both for the internal and external NPMI. As expected, external NPMI is lower than the internal one, as long as it's related to a corpus different from the training one.

	LDA	SCHOLAR	BAT
Internal NPMI	0.1699	0.3203	0.3359
External NPMI	0.00137	0.00870	0.05089

Table 2. NPMI value computed on the internal corpus and externally using Palmetto for LDA, SCHOLAR and BAT

We can look at the Italian version and similarly compute the average NPMI value for each of the generated topics (Table 3). Also in this case the average coherence of BAT outperforms the baseline, holding on average a better coherence with respect to the other groups.

We report on Table 4 and on Table 5, some results regarding the NPMI values for the two different models, the english one and the italian one.

	LDA	SCHOLAR	BAT
NPMI	0.1327	0.2048	0.2512

Table 3. NPMI value compute on the internal Italian corpus for LDA, SCHOLAR and BAT

NPMI	SCHOLAR + BAT
0.5815	israel israeli arab lebanon arabs lebanese palestinian territory jews village
0.5380	encryption encrypt security cryptography enforcement clipper secure key
0.4812	jesus christ lord bible sin church god holy heaven doctrine

Table 4. Computed NPMI on SCHOLAR + BAT for 20 newsgroup

NPMI	SCHOLAR + BAT
0.5518	allegri tuttosport bonucci khedira champions bianconeri marchisio pogba
0.4118	satelliti libici milizie balcani palestina tw merkel laden iraq libia
0.3968	wta tabellone ranking atp finals girone pennetta torneo tennista gironi

Table 5. Computed NPMI on SCHOLAR + BAT for Italian news articles dataset

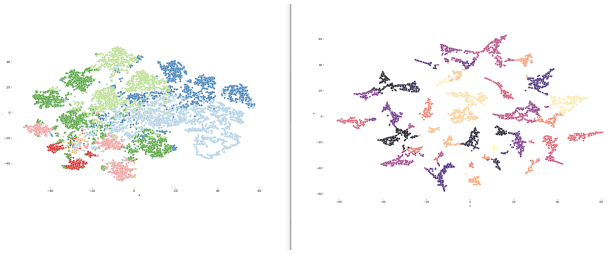

Figure 3. t-SNE on LDA (left) and SCHOLAR+BAT (right) on 20 newsgroup articles over k=50 topics

A visualization based on t-SNE[2] is then proposed on figure 3, this visualization aims at representing all the articles with the probability for each topic using dimensionality reduction over all the topics.

The distribution is more distinct and compact on SCHOLAR + BAT, which holds an overall better performance in terms of coherence.
The same visualization holds a similar result also for the Italian version of SCHOLAR + BAT available in Appendix C 7.3. Here again we can find definitely a better distinction between the topics, this time a little bit noisier than the previous one. On Appendix B 7.2 a stronger comparison in terms of distribution is proposed on SCHOLAR and SCHOLAR + BAT models.

[2]The t-distributed Stochastic Neighbor Embedding is a tool to visualize high dimensional data[6]

5.2. Impact on individual topics

In the previous analysis we showed that BAT holds better coherence on average with respect to LDA and SCHOLAR. A second analysis is now carried out in order to identify topics created by the two models.
Topics are aligned between SCHOLAR and SCHOLAR+BAT using the Jensen-Shannon divergence. It is computed using the similarity on word distribution for each topic. In Appendix B 7.2 are shown some examples of aligned topics. The first table is referred to topics created on 20 newsgroup and the second on Italian dataset. The JS divergence is not strictly comparable between the two tables due to the fact that the first is based on a vocabulary with 1995 words and the second one on a 24496 one. It is noticeable that some topics are almost the same with some small differences. We can take as an example the first one on table 10 the word "warrant" present on SCHOLAR in SCHOLAR+BAT has been replaced by "encryption" which is more harmonic the rest of the topic. Some topics seems to be very different despite of the small JS divergence, this is because the distance is taken on all 100 words that represent the topic.

5.3. Classification of articles

A final analysis on classification of new articles is now proposed. First we have classified twenty new articles from "La Stampa", half about "Sport" and the other half about "Economy". Even if they where not labeled with the same topic results seems to be very promising, since they are able to recognize more specific topics such as "Tennis" and "Volleyball". More information can be find in appendix A and in appendix D 7.1 and appendix D 7.4.
Some other interesting results are shown in Table 6[3] and 7[4], in which we rank the topics with the highest probabilities for two economy articles from the previous analysis. Both of them rank first a topic strictly connected to economy, so they seems pretty correctly classified, but if we look also at the second topic representations, we can highlight that also subtopics of the document are pretty coherent. Specifically in article of Table 6, the topic is focused on the analysis and reports of PIL in Italy, the representation present some words connected to economical analysis. In table 7 as well the second higher distribution is about software and cloud application, in fact in the second article an analysis of a big tech company is represented and this subtopic is correctly interpreted.

[3]First article link: https://www.lastampa.it/economia/2021/07/09/news/istat-l-economia-riparte-ma-l-italia-e-piu-povera-1.40479696

[4]Second article link: https://www.lastampa.it/economia/2021/07/08/news/dopo-il-covid-c-e-la-ripresa-ma-restano-rischi-e-incertezze-1.40475605

Probabilities	**Topic representation**
0.535	redditi comparti gettito inps stipulati contribuenti canone entrate contributi pensionati
0.206	equita dax analisti warning trimestrali ipo qe dollaro bund fed
0.095	vendrame dematerializzazione clienti fornisce software applicazioni ceo cloud applicativi dispositivi

Table 6. First article about economy

Probabilities	**Topic representation**
0.181	redditi, comparti, gettito, inps, stipulati, contribuenti, canone, entrate, contributi, pensionati
0.117	vendrame, dematerializzazione, clienti, fornisce, software, applicazioni, ceo, cloud, applicativi, dispositivi
0.095	infangate, sfollato, egidio, trastevere, gianroberto, grillo, centrodestra, casaleggio, leopardi, ballottaggio

Table 7. Second article about economy

6. Conclusion

During our work, we implemented a new model for the Italian language of BAT, which achieves better results with respect to the other proposed models. The results are definitely encouraging, although no framework is available to systematically test coherence with an external dataset for the Italian language.
Results from the classification on new documents are still rough, but they suggest, that further development and fine tuning of the current model could lead to an improvement in terms of coherence.

We found several further works to propose and some suggestions for further developments. A first idea corresponds to the creation of a recommendation system in the editorial context. An implementation of the model in the editorial context, could really enhance the readers experience, making a huge improvement in the recommendation, which now is based only on more general categories and writers tags. A great problem in this field to focus our study on, is the arise of different new topics, consisting of important words not contained in the vocabulary, for example all the words related to Covid-19 pandemia were not present and so inference on more recent articles is more difficult, due to the massive amount of references to this event. So an interesting and beneficial further developments, would be to explore and compare different and more recent databases and different pre-trained teachers, based on a progressive timeline.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022, 2003.

[2] Dallas Card and Chenhao Tan Noah A.Smith. Neural models for documents with metadata. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1(1):2031–2040, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1(1):4171–4186, 2019.

[4] Alexander Hoyle, Pranav Goel, and Philip Resnik. Improving neural topic models using knowledge distillation. *CoRR*, abs/2010.02377, 2020.

[5] Jon D. McAuliffe and David M. Blei. Supervised topic models. *Proceedings of NIPS*, (1), 2008.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] 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.

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

7. Appendices

7.1. Appendix A.

Id articolo	**Topic 1**	Score Topic 1	**Topic 2**	Score Topic 2	**Topic 2**	Score Topic 2
1	0	0.059	28	0.048	9	0.045
2	24	0.101	47	0.098	16	0.048
3	16	0.157	13	0.114	6	0.075
4	13	0.535	28	0.206	47	0.095
5	13	0.181	47	0.117	49	0.095
6	28	0.088	0	0.047	9	0.038
7	28	0.442	47	0.065	13	0.06
8	47	0.652	16	0.066	13	0.064
9	28	0.256	47	0.027	0	0.026
10	28	0.649	47	0.028	5	0.011
11	3	0.408	17	0.065	2	0.052
12	4	0.818	17	0.124	45	0.002
13	3	0.161	17	0.12	2	0.078
14	45	0.162	3	0.13	17	0.102
15	1	0.548	17	0.114	2	0.057
16	2	0.212	27	0.08	3	0.076
17	3	0.564	17	0.165	2	0.06
18	15	0.094	45	0.091	17	0.061
19	17	0.104	4	0.102	2	0.045
20	1	0.437	2	0.095	27	0.035

Table 8. First 10 articles taken from section economy, last 10 articles from section sport

Id topic	**First ten words belonging to the topic**
0	instagram showgirl cantante sexy attrice film belen conduttrice social cast
1	wta tabellone ranking atp finals girone pennetta torneo tennista gironi
2	diagonale giani palloni compagne volley break mischia palle blengini palla
3	allegri tuttosport bonucci khedira champions bianconeri marchisio pogba chiellini morata
4	decimi dani pedrosa crutchlow iannone phillip ducati hamilton ricciardo jorge
13	redditi comparti gettito inps stipulati contribuenti canone entrate contributi pensionati
15	showgirl instagram sportal belen vales cantante tifosi tuttonapoli conduttrice club
16	imprenditoriale internazionalizzazione confindustria sviluppo export mf pmi bando aise interscambio
17	motogp gp motomondiale jorge valentino malesia sepang marquez lorenzo piloti
24	that sukhoi to que is not transitions signature this alare
28	equita dax analisti warning trimestrali ipo qe dollaro bund fed
45	tifosi tuttonapoli stadio microfoni napoli sportal tifo fiorentina scudetto allenatore
47	vendrame dematerializzazione clienti fornisce software applicazioni ceo cloud applicativi dispositivi

Table 9. Topics that the ”La Stampa” articles are classified to

7.2. Appendix B.

Pair	SCHOLAR vs SCHOLAR+BAT	JS divergence
1	SCHOLAR: escrow, wiretap, crypto, clipper, nsa, warrant, secure, encrypt, key, chip SCHOLAR+BAT:escrow, wiretap, nsa, clipper, crypto, secure, encryption, key, chip, encrypt	0.0485
4	SCHOLAR:jews, nazi, jew, arab, arabs, israeli, islamic, turkey, israel, turks SCHOLAR+BAT:israel, israeli, arab, lebanon, arabs, lebanese, palestinian, territory, jews, village	0.0636
7	SCHOLAR:penalty, puck, goal, score, tie, season, shot, playoff, defensive, ice SCHOLAR+BAT:score, season, rangers, montreal, goal, leafs, penalty, puck, minnesota, playoff	0.0655
15	SCHOLAR:existence, atheist, atheism, belief, universe, exist, faith, science, evidence, truth SCHOLAR+BAT:morality, moral, objective, belief, truth, absolute, christianity, definition, god, faith	0.0783
25	SCHOLAR:helmet, rider, bike, ride, dog, motorcycle, dod, rear, wheel, left SCHOLAR+BAT:bike, ride, car, rider, helmet, honda, engine, dod, dealer, bmw, seat	0.1007
50	SCHOLAR:kevin, roger, gmt, dos, vote, ryan, sea, jeff, thomas, james SCHOLAR+BAT:bike, article, write, surrender, ride, car, gordon, honda, mike, bmw	0.2315

Table 10. Comparison between aligned topics created by SCHOLAR and SCHOLAR+BAT on 20 newsgroup dataset

Pair	SCHOLAR vs SCHOLAR+BAT	JS divergence
1	SCHOLAR: ricciardo, raikkonen, rosberg, kimi, hamilton, vettel, verstappen, sainz, austin, bottas SCHOLAR+BAT:decimi, dani, pedrosa, crutchlow, iannone, phillip, ducati, hamilton, ricciardo, jorge	0.0867
2	SCHOLAR:imbullonati, accise, tasi, gettito, clausole, imu, ires, spending, irap, irpef SCHOLAR+BAT:pensioni, flessibilità, esodati, fornero, damiano, spending, tasi, salvaguardia, precoci, ires	0.111
9	SCHOLAR: esondato, esondazione, veiano, allagate, sannio, allagati, allagamenti, benevento, nubifragio, detriti SCHOLAR+BAT:allagamenti, precipitazioni, temporali, nubifragi, esondazione, piogge, nubifragio, alluvione, esondato, disagi	0.1414
15	SCHOLAR:borussia, marchisio, bianconera, vinovo, khedira, barzagli, bianconeri, siviglia, juve, cuadrado SCHOLAR+BAT:allegri, tuttosport, bonucci, khedira, champions, bianconeri, marchisio, pogba, chiellini, morata	0.1571
25	SCHOLAR:helmet, rider, bike, ride, dog, motorcycle, dod, rear, wheel, left SCHOLAR+BAT:bike, ride, car, rider, helmet, honda, engine, dod, dealer, bmw, seat	0.1856
50	SCHOLAR:intralciata, impedita, iadeluca, erri, sabotata, mandela, monocratico, tav, sabotare, immacolata SCHOLAR+BAT:concussione, incarichi, cassazione, provveditorato, turbativa, indebita, appalto, risarcire, mazzette, arconate	0.2847

Table 11. Comparison between aligned topics created by SCHOLAR and SCHOLAR+BAT on Italian dataset

7.3. Appendix C

Figure 4. t-SNE of SCHOLAR + BAT on italian corpora

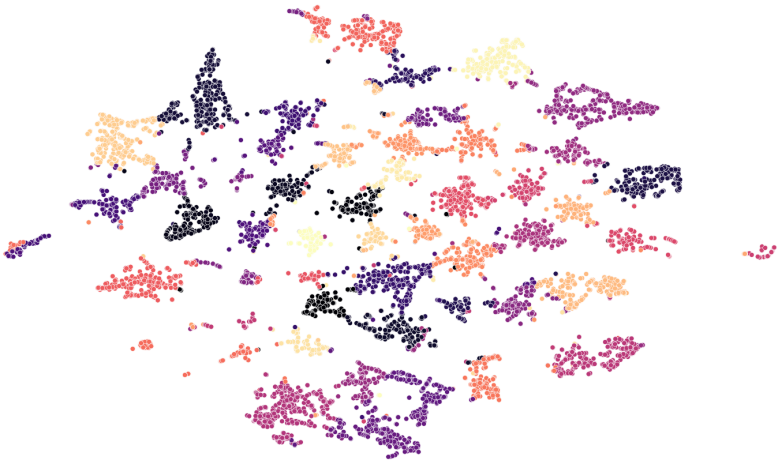

Figure 5. t-SNE of SCHOLAR on italian corpora

Figure 6. t-SNE of LDA on italian corpora

7.4. Appendix D

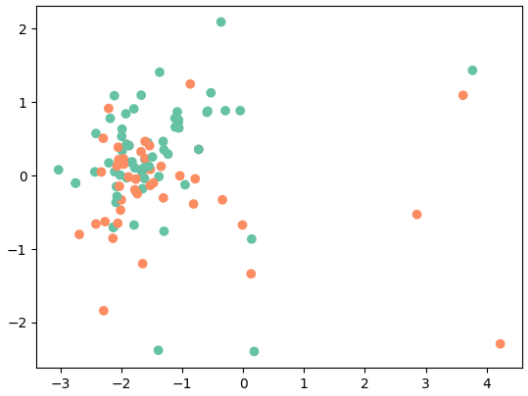

Figure 7. Word embeddings for each word of the classified topics, divided by the section of the original articles

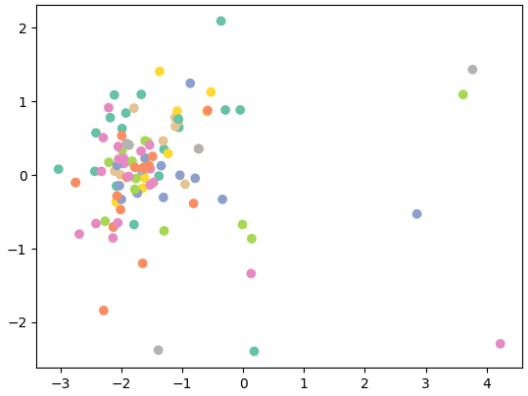

Figure 8. Word embeddings for each word of the classified topics, divided by topics

This two images represent a word embedding of the 20 Italian articles that have been classified by SCHOLAR+BAT (figure 8). It is noticeable how word embedding is not able to separate them properly, even if this should be a good approach. A reason why this does not match the expectation is that we used an Italian embedding. Indeed a lot of words that represent topics are in English or proper names, which are not present into the fastText embedding vocabulary.