Topic Modeling with Transformers for Sentence-Level Using Coronavirus Corpus

https://doi.org/10.3991/ijim.v16i17.33281

Sara Mifrah(✉), El Habib Benlahmar
Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Casablanca, Morocco
mifrah.sara@gmail.com

Abstract—A Topic Model is a class of generative probabilistic models which has gained widespread use in computer science in recent years, especially in the field of text mining and information retrieval. Since it was first proposed, it has received a large amount of attention and general interest among scientists in many research areas. It allows us to discover the mix of hidden or "latent" subjects that differs from one document to another in a given corpus. But since topic modeling usually requires the prior definition of some parameters - above all the number of topics k to be discovered -, model evaluation is decisive to identify an "optimal" set of parameters for the specific data. Latent Dirichlet allocation (LDA) and Bidirectional Encoder Representations from Transformers Topic (BerTopic) are the two most popular topic modeling techniques. LDA uses a probabilistic approach whereas BerTopic uses transformers (BERT embeddings) and class-based TF-IDF to create dense clusters.

Keywords—topic model, sentence-level, machine learning, LDA, BERT, BerTopic

1 Introduction

Topic modeling is an unsupervised machine learning technique that scans a set of documents, detects patterns of words and phrases, and automatically clusters groups of similar words and phrases that best characterize a set of documents.

Topic modeling algorithms generate collections of phrases and words that they think are related, so that we can understand what these relationships mean. It consists of word counting and clustering of similar word patterns to infer topics in unstructured data.

When using a thematic model, we are primarily interested in how well the learned themes match human judgments and help us differentiate ideas. But evaluation of these models has, until recently, been customized and application-specific. Assessments ranged from essentially fully automated assessments to manually designed external assessments. Previous external evaluations used acquired subjects to represent a small, fixed vocabulary and compare this distributive domain with human judgments of commonality [1] [4] [6] [17]. However, these evaluations are manually created and often expensive to perform on industry-specific topics. In contrast, intrinsic metrics have evaluated the amount of information encoded per topic, where confusion is a common example [3], however, [5] found that these essential metrics do not always relate to semantically explicable topics.

Language transformers are a new architecture that was first introduced in the 2017 paper "Attention is all you need". Used primarily in the field of natural language processing (NLP) ,[2] the transformers have rapidly become the model selection for many NLP tasks, since they have surpassed earlier model languages in mainstream performance benches.

This paper is structured as follows: Section 2 presents a small description of the two models. Section 3 defines different coherence measures, our experimentations and results presented in section 4, and finally the conclusions of this paper are provided in Section. 5.

2 Topic modeling

Topic models offer a simple way to analyze large volumes of unlabeled text. A "topic" consists of a group of words that frequently appear together. The data mining [8], automatic language processing [9], computational linguistics [10] information retrieval [11] or mobile learning research [16] communities have studied the extraction or detection of "topics". The approaches dedicated to automatic topic detection originate from several fields: statistics, automatic language processing, linear algebra, etc.

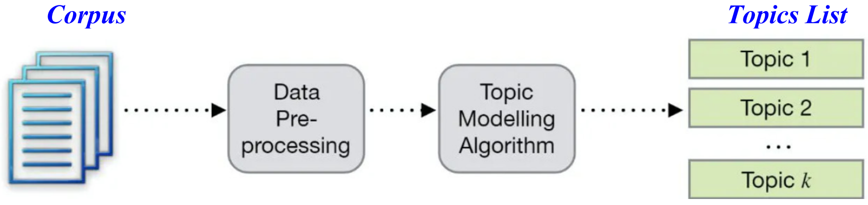


Fig. 1. Topic modeling processing schema

Most topic extraction methods require that the corpus of documents be put in the form of a matrix *A* where the rows represent the documents and the columns represent the words (vector model [12]). Each element *aij* of the matrix contains the weight of the word *wj* in the document *di*, which reflects its importance in the document. A first thematic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998 [13]. Another one, called probabilistic latent semantic analysis (PLSA), was created by Thomas Hofmann in 1999 [14]. Latent Dirichlet Allocation (LDA), perhaps the most widely used thematic model today, is a generalization of PLSA. LDA, developed by David Blei, Andrew Ng, and Michael I Jordan in 2002, introduced Dirichlet prior distributions on document-subject and subject-word distributions, encapsulating the intuition that documents cover a small number of topics and that topics frequently use a smaller number of individual words.

3 Research methodology

3.1 Dataset

In this study, we used a Covid'19 (2019-2020) citation corpus. This corpus is made of papers that were published in the COVID-19 Open Research Dataset (CORD-19) - from https://www.sketchengine.eu/covid19/ -To construct our very own corpus, we select 13214 citations from the Covid'19 corpus, organizing them in a csv file that has the following form "ID; Papers_source; Url; Citations".

Table 1. Extract of the Covid'19 corpus

ID	Paper_source	Url	Citations
1	https://www.biorxiv.org/content/10.1101/2020.01.21.914044v4	biorxiv.org	*Besides most of the reported human-infective Coronaviruses (9) (10) (11) (12) (13) (14) are assigned the lowest p-values predicted by the VHP method (Table S1).*
10	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7182166/	ncbi.nlm.nih.gov	*Betacoronavirus ((β-CoV) Gammacoronavirus (γ-CoV) and Deltacoronavirus (δ-CoV) [1][2]. Since 1960 six different CoVs have been identified and two epidemic CoVs have emerged in humans during the last 2 decades [3]. Severe acute respiratory syndrome [4][5].*
1000	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7033263/	ncbi.nlm.nih.gov	*[1] [2] [3] [4] It was found that the 2019 novel coronavirus (2019-nCoV) was the cause of unexplained viral pneumonia in Wuhan China in December 2019 and was recognized by the World Health Organization (WHO) on January 12 2020.*
1010	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7144984/	ncbi.nlm.nih.gov	*For deeper investigation to find the role that bats play as MERS-CoV risk factor of viral transmission more research is needed [4].*
1017	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7427697 /	ncbi.nlm.nih.gov	*Viral induced cytopathic changes are seen in these cells within 1-2 weeks of infection [2]. However these changes are not specific for NCoV and confirmation using reverse transcription PCR (RT-PCR) is required.*
1019	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3699510/	ncbi.nlm.nih.gov	*In June of 2012 a novel coronavirus designated Middle East respiratory syndrome coronavirus (MERS-CoV) [1] and classified as a 2c betacoronavirus [2] [3] was isolated from a patient with a fatal case of pneumonia and renal failure in Saudi Arabia [3].*

3.2 Pre-processing

A very important task and a critical step in text mining, natural language processing (NLP) and information retrieval (IR) is preprocessing. The pre-processing of data is used to extract non-trivial and interesting knowledge from unstructured textual data.

In preprocessing, the tokenization step is the procedure of dividing a text into words, sentences or other meaningful parts, namely tokens. Tokenization is a kind of segmentation of the text. Generally, segmentation is done by considering only alphabetic or alphanumeric characters that are bounded by non-alphanumeric characters. The Stop Words are those words that are commonly encountered in texts without any particular topic. Another preprocessing step that is used extensively is lowercase conversion. Since upper and lower case forms of words are assumed to be the same, all upper case characters are usually changed to their lower case forms. The last step is uprooting, its purpose is to obtain the root forms of the derived words. Since derived words are semantically similar to their root form, word occurrences are usually calculated after applying stemming on a given text.

3.3 Topic modeling algorithms and coherence measures

LDA. LDA is a mixture model that takes into account the exchangeability of words and documents [9]. The assumption of word exchangeability in a document implies that the order of words in a document is not important, nor is the order of documents in a corpus. The approach of LDA is a hierarchical model. It is assumed to generate each document by the following generative process, where words are generated independently of other words, thus following a unigram bag-of-words model [7]:

To generate a document:

1. Randomly choose a distribution over topics.
2. For each word in a document:
   (a) Randomly choose a topic from the distribution over topics.
   (b) Randomly choose a word from the corresponding topic (distribution over the vocabulary).

More formally, the generative process finds the joint distribution of the hidden and observed variables [7]:

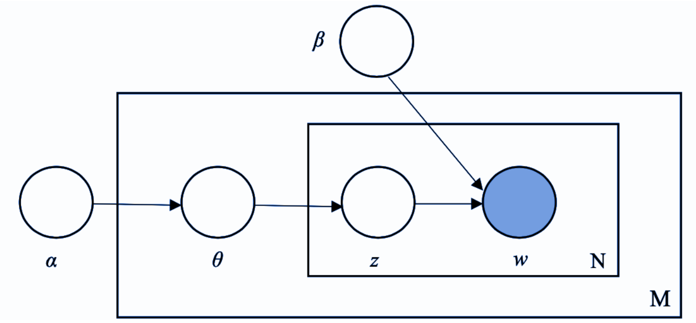


Fig. 2. LDA representation

BerTopic. The BerTopic algorithm is composed of 3 steps [15]:

1. Embedding of textual data (documents).

In this step, the algorithm extracts the embedding from the documents with BERT, or it can use any other embedding technique.

2. Grouped documents

It uses UMAP to reduce the dimensionality of the embeddings and the HDBSCAN technique to group the reduced embeddings and create clusters of semantically relevant papers.

3. Building a thematic representation

The final step is to extract and reduce themes using the class-based TF-IDF technique and then improve word consistency using the maximum marginal relevance study.

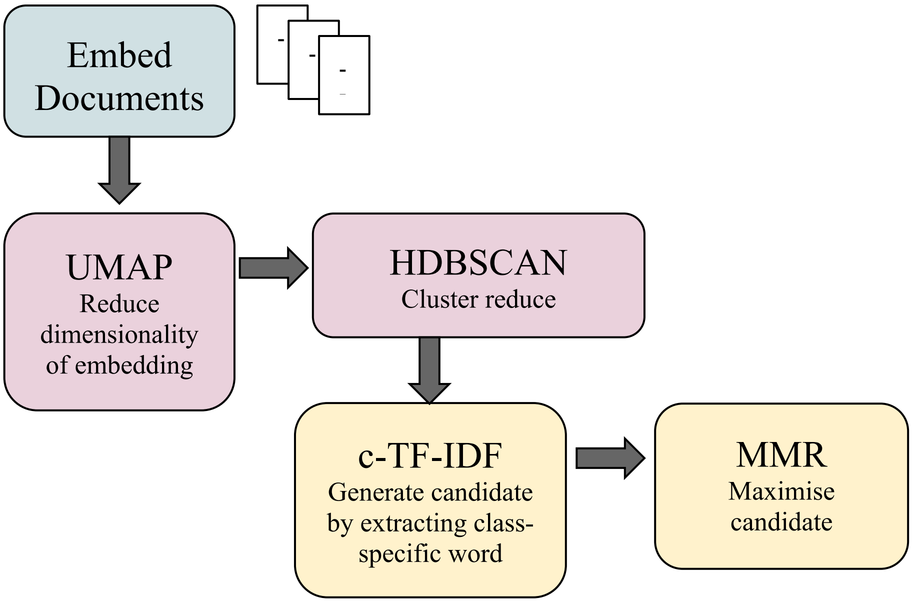

Embed Documents
UMAP
Reduce dimensionality of embedding
HDBSCAN
Cluster reduce
c-TF-IDF
Generate candidate by extracting class-specific word
MMR
Maximise candidate

Fig. 3. BerTopic process

C_v coherence measure. A set of representations or events is said to be coherent if they are mutually supportive. Thus, a coherent set of facts can be interpreted in a context that covers all or most of the facts.

The C_v metric is a sliding window, a segmentation into a set of the most significant words, and an indirect confirming metric that uses standardized mutual information (NPMI) and cosine similarity.

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \mathbf{NPMI}(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|}$$

$$\mathbf{NPMI}(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log(P(w_i, w_j)+\epsilon)} \right)^{\gamma}$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2}$$

4 Experimentation and result

After the preprocessing phase we tried to extract 10 topics for the LDA model, and 8 Topics for BerTopic.

4.1 LDA

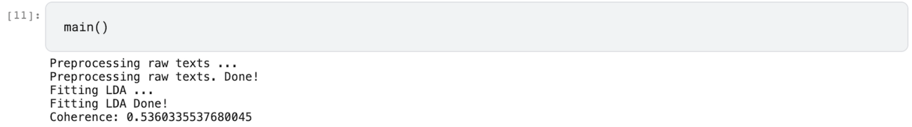
Table 2. LDA topics list

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
protein	transmission	use	cell	sequence
bind	use	test	protein	virus
structure	human	figure	receptor	coronavirus
may	virus	table	ace	bat
fig	evidence	epitope	bind	genome
affinity	model	sequence	entry	share
antibody	estimate	result	target	human
human	base	positive	show	identity
region	animal	detection	virus	identify
analysis	study	sample	also	high

Table 3. LDA topics list

Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
severe	treatment	level	case	patient
case	disease	gene	patient	infection
cause	virus	induce	confirm	study
report	infection	activity	test	infect
infection	control	design	hospital	clinical
spread	vaccine	significantly	suspect	symptom
acute	drug	peptide	positive	covid
respiratory	may	concentration	infection	day
human	effective	construct	virus	report
virus	cause	complex	local	respiratory

After extracting the 10 topics we are calculating the coherences value of LDA model: LDA Coherence: 0.5360335537



```
[11]:  main()

Preprocessing raw texts ...
Preprocessing raw texts. Done!
Fitting LDA ...
Fitting LDA Done!
Coherence: 0.5360335537680045
```

4.2 BERTopic

After generating the c-TF-IDF representations, we have a set of words that describe a collection of documents. Technically, this does not mean that this set of words describes a consistent topic. In practice, we will see that many words describe a similar topic, but some words will somehow overtake the documents.

Topic 0

```
[('china', 0.021690888379580296),
 ('2019', 0.018667881902309315),
 ('wuhan', 0.016660202348458265),
 ('december', 0.015437453690974778),
 ('coronavirus', 0.014363793485995759),
 ('disease', 0.012582911912390218),
 ('novel', 0.011165874080245549),
 ('acute', 0.010987124429209134),
 ('2020', 0.010928756842227372),
 ('outbreak', 0.009943952109810722)]
```

Topic 1

```
[('epitopes', 0.0385428524429711),
 ('vaccine', 0.02887959865572832),
 ('cell', 0.021905552150179045),
 ('mhc', 0.018388655032853595),
 ('vaccines', 0.016768573612687448),
 ('identical', 0.013373017202347789),
 ('derived', 0.012973151951999063),
 ('table', 0.012360157283793671),
 ('development', 0.012067216046675489),
 ('epitope', 0.010730059999849646)]
```

Topic 2

```
[('middle', 0.026223285256662347),
 ('east', 0.02607180012047407),
 ('2012', 0.02356526823900479),
 ('saudi', 0.02316906769534301),
 ('arabia', 0.022439228863824999),
 ('mers', 0.015912430894817697),
 ('syndrome', 0.015256936262711318),
 ('september', 0.01454399436701444),
 ('respiratory', 0.013870139196066581),
 ('first', 0.012523978687433952)]
```

Topic 3

```
[('origin', 0.01003242131700905),
 ('continuously', 0.008649615606785021),
 ('sars', 0.008490716960446723),
 ('category', 0.008289284146047023),
 ('mlv', 0.008167186514344456),
 ('ci', 0.00799783959942518),
 ('gap', 0.007902608820231095),
 ('cov', 0.007878621657518254),
 ('95', 0.007647592254182505),
 ('field', 0.007548767069062219)]
```

Topic 4

```
[('preprint', 0.061305605168092005),
 ('copyright', 0.04547395432865204),
 ('holder', 0.045209570873252905),
 ('https10', 0.04412416222439733),
 ('11012020', 0.043757103620619596),
 ('doi', 0.04354369594817463),
 ('peer', 0.04319618810028289),
 ('reviewed', 0.04223833672376924),
 ('03', 0.03872475797849065),
 ('biorxiv', 0.02829241918269239)]
```

Topic 5

```
[('mers', 0.018101242969112905),
 ('2012', 0.016538183766286033),
 ('saudi', 0.014446996530450881),
 ('arabia', 0.013189333830855668),
 ('june', 0.011399747785779523),
 ('reported', 0.010684022186887998),
 ('first', 0.010275514882021696),
 ('cases', 0.01013491936287263),
 ('failure', 0.009284779528442626),
 ('died', 0.008932670148082906)]
```

Topic 6

```
[('ace2', 0.02688160250886786),
 ('rbd', 0.025583383399618992),
 ('binding', 0.02382217371833988),
 ('receptor', 0.01713783839947591),
 ('protein', 0.013599907594735568),
 ('spike', 0.013140605032614576),
 ('affinity', 0.012266665675595442),
 ('amino', 0.011649926573456157),
 ('rbm', 0.010874417747854869),
 ('acids', 0.009697292049809743)]
```

Topic 7

```
[('aerosol', 0.030737821695766023),
 ('droplets', 0.028031644232077962),
 ('airborne', 0.02586227774525122),
 ('transmission', 0.019351740742362906),
 ('contact', 0.017589746212067114),
 ('surfaces', 0.015190111009437844),
 ('through', 0.013429164732121628),
 ('aerosols', 0.012392831791803030),
 ('via', 0.011192074875565757),
 ('person', 0.010741287194714198)]
```

To improve word consistency, maximum marginal relevance was used to find the most consistent words without having too much overlap between the words themselves. This eliminates words that do not contribute to a topic.

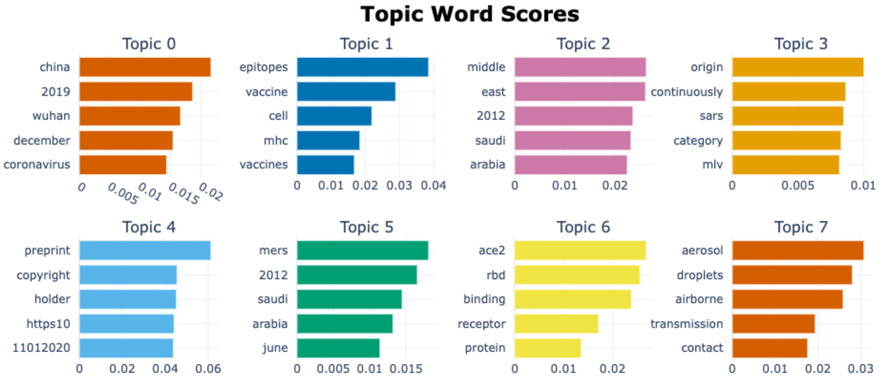


Fig. 4. Topic word scores

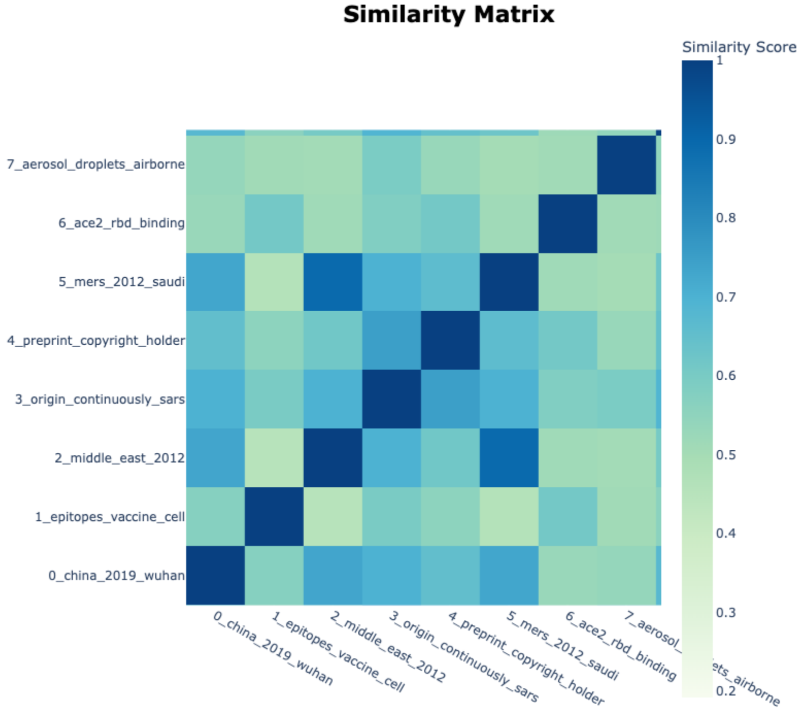


Fig. 5. BerTopic similarity matrix

5 Conclusion

This paper presents how modern language transformers can be used in topic modeling. Building on traditional topic modeling algorithms, transformers can optimize the text vectorization as well as the final synthesis output. In practice, topic modeling is always a difficult task because finding topics in large documents is never easy for a human. But to further incorporate deep learning into topic modeling, we will still have to wait for some research breakthroughs.

Currently, human intervention is unavoidable in topic modeling because some manual checks are still needed in the evaluation phase.

6 References

[1] Jurgens, D., & Stevens, K. (2010) ‘The S-Space package: An open source package for word space models’. Proceedings of the ACL 2010 systems demonstrations, Uppsala, Sweden.

[2] Vaswani, Ashish, et al. (2017) “Attention is all you need.”, Advances in neural information processing systems.

[3] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno (2009)” Evaluation methods for topic models”. In Proceedings of the 26th International Conference on Machine Learning (ICML). Omnipress. https://doi.org/10.1145/1553374.1553515

[4] Sara Mifrah and El Habib Benlahmar. (2019) ‘Semantic Relationship Study between Citing and Cited Scientific Articles Using Topic Modeling’. In Proceedings of the 4th International Conference on Big Data and Internet of Things (BDIoT’19), Rabat Morocco. 2019. https://doi.org/10.1145/3372938.3372943

[5] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber and D. M. Blei (2009) "Reading tea leaves: How humans interpret topic models", Proc. Adv. Neural Inf. Process. Syst., pp. 288-296.

[6] Hourrane, Oumaima, et al. (2018) "Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers." International Conference on Big Data, Cloud and Applications. Springer, Cham 2018. https://doi.org/10.1007/978-3-319-96292-4_15

[7] S. Clark (2013). Topic Modeling and Latent Dirichlet Allocation. Available at: https://www.cl.cam.ac.uk/teaching/1213/L101/clark_le ctures/lect7.pdf

[8] Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R (2008): A new document clustering algorithm for topic discovering and labeling. Progress in Pattern Recognition, Image Analysis and Applications, 161-168. https://doi.org/10.1007/978-3-540-85920-8_20

[9] Blei, D., Ng, A., & Jordan, M. (2003): Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022.

[10] Ferret, O. (2006): Approches endogènes et exogènes pour améliorer la segmentation thématique de documents. Traitement Automatique des Langues 47, 111– 135.

[11] Zamir, O., Etzioni, O., & Madani, O. (1997): Fast and intuitive clustering of web documents. KDD'97, 287–290.

[12] Salton, G., Wong, A., & Yang, C. (1975): A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620.14. https://doi.org/10.1145/361219.361220

[13] Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, San- tosh (1998). "Latent Semantic Indexing: A probabilistic analysis"(Post- script). Proceedings of ACM PODS: 159– 168. ISBN 978-0897919968. https://doi.org/10.1145/275487.275505

[14] Hofmann, Thomas (1999):"Probabilistic Latent Semantic Indexing" (PDF). Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. Archived from the original (PDF) on 2010-12-14. https://doi.org/10.1145/312624.312649

[15] Grootendorst, M.R. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv, abs/2203.05794.

[16] Hamzah, A., Hidayatullah, A. F., & Persada, A. G. (2020). Discovering Trends of Mobile Learning Research Using Topic Modeling Approach. International Journal of Interactive Mobile Technologies (iJIM), 14(09), pp. 4–14. https://doi.org/10.3991/ijim.v14i09.11069

[17] Chanaa, A., & El Faddouli, N.- eddine. (2021). E-learning Text Sentiment Classification Using Hierarchical Attention Network (HAN). International Journal of Emerging Technologies in Learning (iJET), 16(13), pp. 157–167. https://doi.org/10.3991/ijet.v16i13.22579

7 Authors

Sara Mifrah is with Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben M’sik, Casablanca, Morocco (email: mifrah.sara@gmail.com).

El Habib Benlahmar is with Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben M’sik, Casablanca, Morocco (email: h.benlahmer@gmail.com).

Article submitted 2022-06-13. Resubmitted 2022-08-01. Final acceptance 2022-08-05. Final version published as submitted by the authors.