International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.14, No.1 (Jul-23)

http://dx.doi.org/10.12785/ijcds/140101

AraBERTopic: A Neural Topic Modeling Approach for News Extraction from Arabic Facebook Pages using Pre-trained BERT Transformer Model

Nassera HABBAT[1], **Houda ANOUN**[1] **and Larbi HASSOUNI**[1]

[1]*RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco*

Received 21 Jun. 2022, Revised 6 May. 2023, Accepted 7 May. 2023, Published 1 Jul. 2023

Abstract: Topic modeling algorithms can better understand data by extracting meaningful words from text collection, but the results are often inconsistent, and consequently difficult to interpret. Enrich the model with more contextual knowledge can improve coherence. Recently, neural topic models have emerged, and the development of neural models, in general, was pushed by BERT-based representations. We propose in this paper, a model named AraBERTopic to extract news from Facebook pages. Our model combines the Pre-training BERT transformer model for the Arabic language (AraBERT) and neural topic model ProdLDA. Thus, compared with the standard LDA, pre-trained BERT sentence embeddings produce more meaningful and coherent topics using different embedding models. Results show that our AraBERTopic model gives 0.579 in topic coherence.

Keywords: Neural topic model, ProdLDA, AraBERT, Topic coherence.

1. Introduction

Nowadays, because of the exponential development of the Internet, a huge quantity of documents, such as online news are produced every day, especially in social media like Facebook which is one of the most important social networks in Africa. Concerning Morocco, there were 22 010 000 Facebook users in January 2021 [1]. Mining the knowledge and topics from social media posts have attracted a lot of attention these last years. To discover hidden topical patterns which are present in large collections of texts, many unsupervised techniques are usually used to generate probabilistic topic models. Among these models, we find Non-negative Matrix Factorization (NMF), Latent Semantic Indexing (LSI), and Latent Dirichlet Allocation (LDA) [2] .

However, those probabilistic models do not take advantage of language model pre-training benefits. Many extensions were proposed to integrate different types of information and add contextual knowledge to the topic models. The most prominent architecture in this category is Bidirectional Encoder Representations from Transformers (BERT) which allows us to extract representations from pre-trained documents to easily reach state-of-the-art performance through numerous tasks [3].

Recently, some neural topic models were explored and have shown promising results. For instance, ProdLDA, which is a developed version of LDA based on deep learning, is an expert products instead of mixture model in LDA to yield much more interpretable topics [4].

The main contribution of this paper is to present a proposed model (AraBERTopic) to extract topics from Arabic news published on Facebook pages using AraBERT as language model pre-training and ProdLDA as a neural topic model. To prove the high performance of our model; we compared, on the one hand, its feature extraction phase with different embedding models (Glove, Doc2Vec, and Asafaya as a bert-base-arabic model [5]), and we compared, on the other hand, its topic model phase with standard LDA. The results demonstrate that our proposed model is superior to other models in terms of Normalized topic coherence, Pointwise Mutual Information (NPMI), and perplexity metrics. The rest of this paper is organized as follows: Section 2 provides a brief review of literature. The proposed model is described in 3rd Section. In section 4, we present the results and discusses. Finally, the paper is summarized and the future work has prospected.

2. Related Works

We introduce in this part some topic extraction methods, including traditional feature extraction methods and deep learning methods.

Probabilistic topic models are much used in natural

E-mail address: nassera.habbat@gmail.com, houda.anoun@gmail.com, lhassouni@hotmail.com language processing (NLP), and among the most commonly used methods, we find LDA. In [6]; the author reviews the academic papers on LDA topic modeling published from 2003 to 2016. In [7], the authors examine the feedback on Duolingo, a free and enjoyable language-learning program. They employed LDA to figure out the points that consumers bring up and to create a rough outline for both software developers and people looking for apps that are comparable based on these points. Similarly, the authors in [8] examined the content of startups using LDA to see how their communication approaches can change as they scale. They found that the contents may be categorized into five different topics using data from Twitter as a source of information.

To imitate the statistical process of LDA, the authors in [9] investigate the possibility of using deep neural network to model the statistical process to minimize the computational time in LDA; Therefore, they proposed two deep neural network variants: two and three Neural Network (NN) DeepLDA, in their experiments they used Reuters-21578 as a dataset, and some standard libraries in Python like genism, NLTK and Keras and to record the accuracy of the models, a Support Vector Classifier (SVC) was used. Their results showed that 3NN DeepLDA outperforms 2NN DeepLDA and LDA.

In recent years, deep learning has become a powerful machine learning technology, which allows learning multi-level representation and several methods exist that are particularly adapted for learning meaningful hidden representations. Among those models, we find the Variational Autoencoder (VAE) which is a deep generative model. In [4] the authors present the first efficient autoencoding variational Bayes (AEVB) which is an inference method based on latent Dirichlet allocation (LDA)-(AVITM), in this paper, they proposed a novel topic model named ProdLDA which replace the hybrid model in LDA with expert products, After applying their model on 20 Newsgroup dataset they obtained that AVITM outperforms baseline methods in term of accuracy and reasoning time, and the topics given by ProdLDA are more explanatory, Similarly, the authors in [10] presented Neural Variational Correlated Topic Model consisting of two main parts; the 1st one is the inference network with Centralized Transformation Flow and the 2nd one is the multinomial softmax generative model. To evaluate their model they used NPMI topic coherence. Their results showed that the model enhances the performance of topic modeling and can effectively capture topic correlation. However, the authors in [11] used LDA to identify the topics of discussion in Tweets, resulting in a directed multilayer network in which users (in one layer) are linked to discussions and topics (in a second layer) in which they contributed, with interlayer connections indicating user participation in discussions. Although there are other methods for topic modeling on short texts, such as the Biterm Topic Model. The authors in [12] used this technique with part-of-speech tagging on noun-only to analyze the public mapping review as a data source regarding hospitals in order to identify the topics in the review with a low score so that it could be used as a suggestion for enhancing health services.

Only some recent studies have used semantic embeddings like Bidirectional Encoder Representations from Transformers (BERT) and ELMO (Embeddings from Language Models) in topic analysis. In [13] the authors proposed Variational Auto-Encoder Topic Model (VAETM) which combines entity vector representation and word vector representation. The model uses large-scale external corpus and manually edited large-scale knowledge graph to learn the embedding representation of each word and entity, then, those embedding representations are integrated into the VAE framework to deduct the hidden representation of topic distributions; To prove the performance of their model they compared it with various of baseline algorithms (LDA, Sparse Additive Generative Model (SAGE) [14] and SCHOLAR [15]), and 20Newsgroups. IMDB and Chinese Standard Literature (96,000 national and industry standards of China) using as datasets; based on perplexity, NPMI, and accuracy measures; they showed that the model better mine the hidden semantic of short texts and improve topic modeling.

Many researchers opt for BERT [3] as contextualized word representations because it progresses the state of the art for different NLP tasks by pushing the MultiNLI accuracy up to 86%, score to 80.5% for the General Language Understanding Evaluation (GLUE), the Stanford Question Answering Dataset (SQuAD v1.1) question answering Test F1-score to 93.2 and Test F1-score to 83.1 for SQuAD v2.0, compared to Glove, ELMOs, and OpenAI GPT; Among NLP tasks we found Topic modeling. The authors in [16] include in a neural topic model, the contextualized BERT embeddings to get more consistent topics compared to Neural-ProdLDA, NVDM, and LDA.

Moreover, BERT creates word embeddings in multiple languages, in [17] the authors used LDA topic model and multilingual pre-trained BERT embeddings to analyze the evolution of topics in Chinese, English, and multilingual in scientific publications using Google-pre-trained BERT models: “bert-base-chinese” for Chinese, “bert-base-uncased” for English and ¨bert-base-multilingual-uncased” for multilingual text. The results showed that the model can well analyze the scientific evolution of similar relationships between monolingual and multilingual disciplines. In most cases, 80% of the relationships are related to the key topics of each language.

Compared with English, Arabic is a language with rich forms, less syntactic exploration, and fewer resources. The pre-trained AraBERT model [18] is very effective in language understanding compared with multilingual BERT. It achieves the most advanced performance in most Arabic NLP tasks.

3. Proposed approach

We will describe in this part, our model (AraBERTopic), the global architecture is exposed in Figure 1 . There are four principal components:

(1) Data Acquisition: This component aims to collect data from Facebook pages using web scraping;

(2) Pre-training model using AraBERT and extracting the features.

(3) Extraction of topics using Neural Prod-LDA

(4) Evaluation of model comparing it with baselines models.

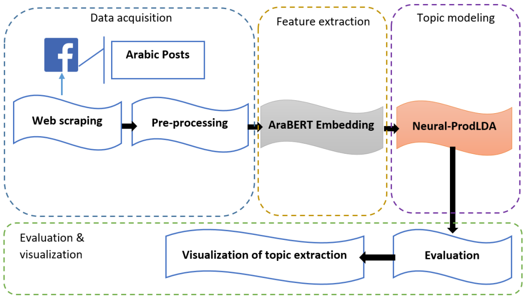


Figure 1. The global architecture of AraBERTopic model.

A. Word Embedding

AraBERT [18] is a pre-training BERT transformer model for the Arabic language. It uses Transformer to learn the context between words (or sub-words) in texts. The Transformer encoder is considered to be bidirectional, which means that it reads the whole word sequence at one time, instead of reading the text input-oriented model in order (from right to left or from left to right).

The input is a token sequence, which is first embedded into a vector and then processed in the neural network to obtain a vector sequence as output. As shown in Figure 2, [CLS] is a special symbol added before each sample input and [SEP] is a special separator mark. Generally, the embedding models include the following three layers:

• The token embeddings layer changes each word tag into a 768-dimensional vector representation.

• The segment embeddings layer has two representations Vector: the 1st vector (index 0) is attributed to all tokens of input 1, and the last vector (index 1) is attributed to all tokens of input 2.

• The Position embeddings layer: AraBERT is designed to process up to 512 input sequences. Therefore, AraBERT must learn a vector representation of each position. This means that the position integration layer is a size look-up table (512, 768) where :

- The 1st row is the representation of the vector of each word in position 1.

- The 2nd row is the representation of the vector of each word in position 2, etc.

These representations are added in a single representation by the elements in the generated form (1, n, 768). The following Figure shows the input representation passed to the AraBERT encoder layer.

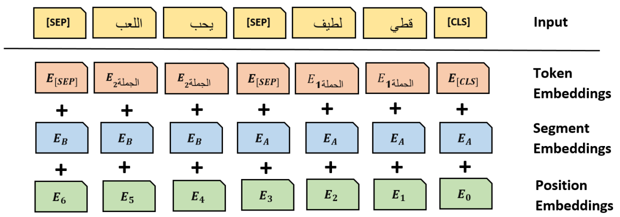


Figure 2. AraBERT input representation.

The learning of AraBERT takes place in two phases: The 1st one is the pre-training, it is done only once, it allows the creation of a neural network that has a certain general understanding of the language. Then, the 2nd phase is called the fine-tuning phase, which allows the network to be trained on a specific task like classification, question answering, and Topic modeling as shown in Figure 3.

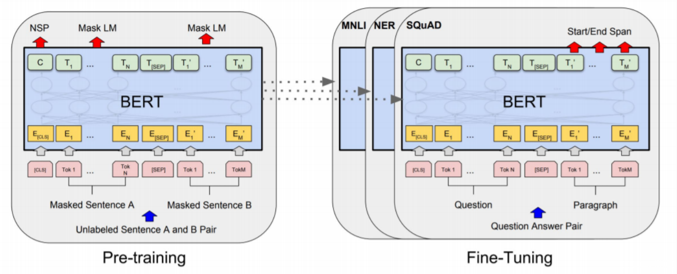


Figure 3. General pre-training and fine-tuning processes for BERT [3].

Concerning Pre-training Dataset, it was scraped from Arabic news websites for articles. In addition there are two major publicly accessible Arabic corpora:

- Open Source International Arabic News (OSLAN) Corpus [18] composed of 3.5 million articles (about 1 billion tokens) from 31 news sources in 24 Arab countries,

- An Arabic corpus of 1.5 billion word [17], which is a contemporary corpus composed of more than 5 million articles from ten major information sources in 8 countries,

After deleting the repeated sentences, the final size of the pre-training dataset in AraBERT is 70 million phrases (about 24 GB), which can represent a large number of topics discussed in news from different Arab regions. AraBERT was assessed on three tasks concerning Arabic language understanding: Named Entity Recognition, Sentiment Analysis, and Question Answering [12]. In our work, we used AraBERT in the Topic modeling task to enrich the representations and provide a significant augmentation in topic coherence by adding to neural topic models, the contextual information.

B. Neural Topic Model

ProdLDA [4] solved the problem of $p(w|\phi, \gamma)$ (a mixture of multinomials) distribution of LDA, which consists of never making predictions that are more precise than the mixed components. This led to low-quality subjects and people's judgment is not consistent. The way to resolve this problem is by using the weighted product of experts to convert mixed words into word level. According to the definition, the weighted product of experts can make clearer predictions than any combination of experts. ProdLDA uses the weighted product of experts to replace the mixed word hypothesis in LDA [19], which greatly improves the consistency of topics.

ProdLDA employed a VAE for LDA using a Laplace approximation for the Dirichlet distribution, which makes it possible to train a Dirichlet variational autoencoder. Moreover, this model does not directly reparametrize the Dirichlet distribution.

To successfully apply VAE to LDA, an encoder network is used that approximates the Dirichlet prior $p(w|\phi, \gamma)$ with a softmax-normal distribution LN. In other words:

$$p(\phi, \gamma) p(\phi|\mu, \Sigma) = LN(\phi|\mu, \Sigma) \tag{1}$$

Where μ and Σ are the encoder network outputs. In addition, the Adam optimizer, batch normalization, and dropout units in the encoder network are used to component collapse. The only modifications to pass from LDA to ProdLDA are that ProdLDA is not normalized, and the conditional distribution of w_m is interpreted as $w_m|\gamma, \phi \sim Multinomial(1, \sigma(\gamma\phi))$.

For the multinomial, the relation to an expert products is very simple; if we take 2 N-dimensional multinomials parameterized by mean vectors p and q and set the natural parameters to $p = \sigma(v)$ and $q = \sigma(r)$, we can show that (where $\delta \in [0, 1]$):

$$P\,(x|\delta \mathbf{v} + (1-\delta)\,\mathbf{r}) = \alpha \prod_{i=1}^{N} \sigma(\delta \mathbf{v}_i + (1-\delta)\,\mathbf{r}_i\ )^{x_i} \ \prod_{i=1}^{N} [v_i^{\delta} .\ r_i^{(1-\delta)}]^{x_i} \tag{2}$$

Where the notation $\alpha(\beta)$ means applying the softmax function separately to each column of the matrix β and δ represents the output of each network is a vector in R^K.

In brief; the used approach trains an encoding neural network to map pre-trained contextualized word embeddings (AraBERT) to latent representations which are variably sampled from a Gaussian distribution and transmitted to a network of decoders. This network of decoders must rebuild the bag-of-words representation of the document.

4. EXPERIMENTS AND RESULTS

We will describe in this part, our dataset (collect and preprocessing), then we will present baselines and used metrics to compare our model with baseline methods.

A. Dataset

For almost 10 years, the Application Programming Interface (API) provided by Facebook has been the primary tool for researchers to collect data on Facebook. These data contain public information about user profiles, comments and reactions to public messages. However, after the Cambridge Analytics (CA) scandal in early 2018, Facebook significantly tightened access to its API [20].

To pull data from Facebook pages, we used web scraping techniques with Python language (Version 3.8) namely Requests (Version 2.25.0) and BeautifulSoup4 (Version 4.9.3) as external libraries for automatic browsing. In our study, we were interested in Arabic posts published on Facebook by Moroccan news pages. We chose seven Facebook pages (Hespress, Medi1TV, aljarida24.ma, alakhbar.maroc, Alyaoum24, JARIDATACHCHAAB, al3omk), and collected 81 598 posts published from 04 October 2020 to 05 March 2021 (details of collected posts are shown in Table I)

TABLE I. DATA COLLECTION

Facebook pages	**Number of collected posts**
Hespress	22 854
Medi1TV	18 176
aljarida24.ma	8 133
alakhbar.maroc	6 400
Alyaoum24	15 489
JARIDATACHCHAAB	1 571
al3omk	8 975
Total	**81 598**

B. Data preprocessing

We performed the common pre-processing steps in existing approaches which consist of:

- Removal of Arabic stopwords.

- Removing hyperlinks, hashes to keep only Arabic text.

- Lemmatization of words using Farasapy Lemmatizer [21] for Arabic text.

- Selecting the most frequent 2,000 words as the vocabulary

C. Word embedding

It designates a set of learning methods in NLP where vocabulary words (or phrases) are converted to numerical vectors. In our research, we used BERT word embeddings. AraBERT is an Arabic pre-trained language model that gives a contextual embeddings, which is used to generate word embeddings. Each word is interpreted as a vector of size 768 and consequently each sentence is a list of word embeddings are extracted. The sequence of the embeddings is completed so that they have the same size. For our AraBERTopic model, we used the pretrained bert-base-arabert Arabic embedding with 12 encoder blocks/layer, 768 hidden dimensions, 12 attention heads, and 110 M parameters. It can be found on the Google Bert model website [22].

D. LDA and ProdLDA settings

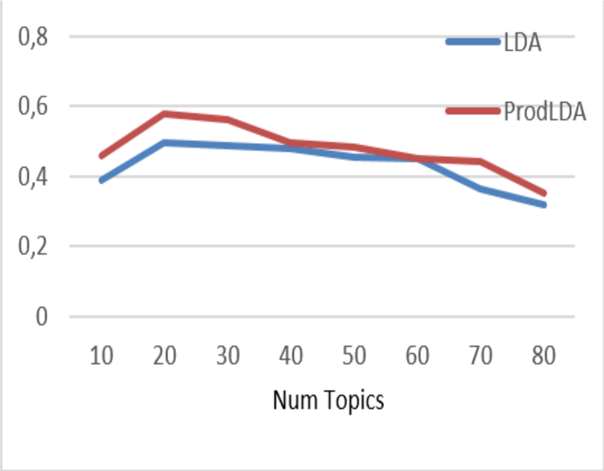


Figure 4. Choosing the optimal number of LDA and ProdLDA topics.

Concerning LDA and prodLDA parameters, we selected twenty topics and the top-ten words for each topic. For the number of topics (N), we tested N values from 5 to 80. When the N value was 20, the results were good with the highest topic coherence value as shown in Figure 4.

E. Experimental Setup

In our experiments, we used the implementation of AraBERTopic in Pytorch Library (Version 1.6.0). Adam optimizer was used, with a batch size of 64 and 2e-3 for the learning rate. Our model was fine-tuned after 10 epochs over the data.

F. Baselines

We compared the two levels of our approach (feature extraction and topic modeling) to other models to prove its performance. Concerning feature extraction level, we compared AraBERT Embedding with three embedding models:

- asafaya/bert-base-arabic model [5]: This model is contextual, it was pretrained on 8.2 Billion words. Glove [23]: The model gives the vector representation of words using an unsupervised learning method. We choose 60% of the data set to train the model, because it captures the overall meaning of sentences in a relatively small memory.

- Doc2vec [24]: This model converts sentences or paragraphs to numeric vectors. In our work, we used Doc2vec Gensim implementation. To train our Doc2vec model, we used the same dataset as for the Glove model. Concerning the second level of topic modeling, we compared ProdLDA to LDA used with different word embedding models already mentioned.

G. Metrics

To assess the performance of the topic model is usually using the following metrics:

1. Subject coherence based on NPMI algorithm,

2. Topic coherence measure,

3. Confusion or Perplexity evaluation.

1) NPMI

This metric [25] gives an automatic measure of the quality of the topics to evaluate our proposed model as well as baselines. It derives from Pointwise Mutual Information (PMI) and measures the effect of one $x_m$variable on another x_n . Its formal definition is as follows:

$$PMI(x_m, x_n) = log\frac{p(x_m, x_n)}{p(x_m)p(x_n)} \tag{3}$$

$p(x_n)$: The likelihood that the word x_n appearing in the corpus,

$p(x_m, x_n)$: The likelihood that the word x_m and word x_n appear together in the corpus.

We took in our experiment the top-five words of each topic, and for each word; we compute the NPMI score following this equation:

$$NPMI = \sum_{m=1}^{j} \sum_{n=m+1}^{j} \frac{PMI(x_m, x_n)}{-\log P(x_m, x_n)} \tag{4}$$

The topics that scored higher in NPMI are the most likely words to seem more often in the same document m than those who occasionally appeared.

2) Topic coherence

Topic coherence measure is also derived from PMI, which is used to evaluate the semantic similarity between high-resolution words of a topic. Topic coherence score is computed as follows:

$$Score_{UCI}(x_i, x_j) = log\frac{p(x_i, x_j) + \varepsilon}{p(x_i)p(x_j)} \quad (5)$$

3) Perplexity

Perplexity (PPL) is a statistical measure used to assess the quality of model subject modeling. It is computed as follows:

$$Perplexity\,(w|z,\theta,\beta) = exp(\frac{-\sum_{m=1}^{M}\sum_{n=1}^{N_m} logp(w_{mn}|z_{mn},\theta_m,\beta)}{\sum_{m=1}^{M} N_m}) \quad (6)$$

Where: N_m : The number of words in the document M.

θ : Document-topic density

β : Topic-word density.

H. Results analysis

In our experiments, we used experimental parameter summarized in the following Table II:

TABLE II. SETTINGS OF OUR IMPLEMENTED MODEL

Parameter	Value
N Components	20
Topic Prior Mean	0.0
Topic Prior Variance	0.95
Batch size	64
Num_epochs	10
Hidden Sizes	(100, 100)
Activation	softplus
Solver or optimizer	Adam
Dropout	0.2
Learning Rate	0.002
Momentum (momentum to use for training)	0.99
Reduce On Plateau (reduce learning rate by 10x on plateau of 10 epochs)	False

Those parameters are used to implement our AraBERTopic using contextualized-topic-models Python library [16].

1) Quantitative Evaluation

We calculated NPMI, topic coherence, and perplexity measure of each model to better compare their performance with different embedding models on short Arabic text. As can be seen in Figure 5, the NPMI of AraBERTopic (AraBERT + ProdLDA) is 0.553, which is higher than the other models. Overall, our model outperforms baselines methods in terms of topic coherence value, perplexity score as shown in Table III.

2) Qualitative evaluation

To prove the quality of the topics extracted by our models using ProdLDA which has the highest score within different embedding models as shown in Table III, we display, in Table IV, examples of the top-six words of three topics from all the models. We have added the translation of each word in English.

TABLE III. EVALUATION OF PERFORMANCE OF TOPIC MODELS PRODLDA AND LDA WITH DIFFERENT EMBEDDING MODELS.

Models		**Metrics**		
Word Embedding	**Topic Model**	**NPMI**	**CV**	**PPL**
AraBERT	**ProdLDA**	0.553	0.579	11.25
	LDA	0.137	0.497	20.65
Asafaya	**ProdLDA**	0.484	0.543	56.20
	LDA	0.128	0.494	61.28
Doc2Vec	**ProdLDA**	0.358	0.482	63.79
	LDA	-0.16	0.477	78.63
Glove	**ProdLDA**	0.333	0.534	86.9
	LDA	0.109	0.434	90

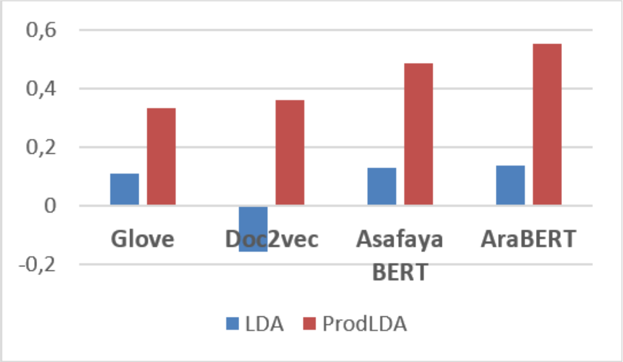
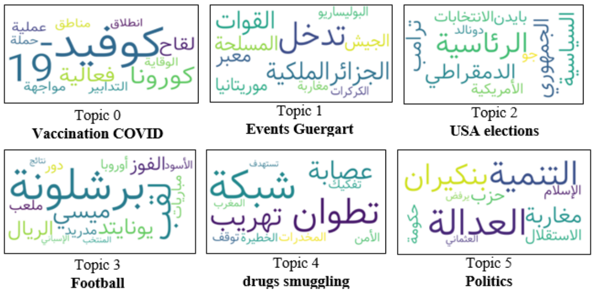


Figure 5. NPMI of different models.

The topics generated using AraBERTopic are more coherent than those generated by other models. In the second position, we find Asafaya's topics which sound to be coherent, but are influenced by additional mixed topics; for example, the 1st topic is about Coronavirus but it includes the term 'Geographic' which is a bit far from the topic.

TABLE IV. TOP SIX-WORDS OF THREE TOPICS EXTRACTED BY ALL THE MODELS..

Embeddings models using ProdLDA	Topics
Glove	['لقاح', 'صلاة', 'الوقاية', 'تفرض', 'كورونا', 'مناطق'] [Places, Corona, Impose, prevention, Prayer, Vaccine] ['الكركرات', 'ترامب', 'الرئيس', 'يضع', 'الملك', 'يوم'] [Day, The king, Put, President, Trump, Guergarat] ['الأمازيغية', 'السياسية', 'النواب', 'بايدن', 'أمريكا', 'حملة'] [Campaign, America, Biden, Representatives, Political, Amazigh]
Doc2vec	['الحجر', 'حملة', 'كورونا', 'الوقاية', 'المواطنين', 'سلطات'] [Authorities, Citizens, Quarantine, Campaign, Corona, prevention] ['الكركرات', 'الإمارات', 'فتح', 'علاقات', 'الملكية', 'أمريكا'] [America, property, Relation, Open, Emirates, Guergarat] ['ترامب', 'النواب', 'بايدن', 'أمريكا', 'قانون', 'مشروع'] [Bill, Law, America, Biden, Representatives, Trump]
Asafaya	['كورونا', 'الصيني', 'بريطانيا', 'لقاح', 'منظمة', 'الجغرافي'] [Geographic, Organization, Vaccine, Britain, Chinese, Corona] ['الجزائر', 'الملكية', 'القوات', 'المسلحة', 'معبر', 'موريتانيا'] [Mauritania, crossing, armed, forces, royalism, Algeria] ['الرئاسية', 'ترامب', 'الانتخابات', 'بايدن', 'الأمريكية', 'جو'] [Joe, American, Biden, Elections, Trump, Presidential]
AraBERT	['حملة', 'لقاح', 'كورونا', 'فعالية', 'العالمية', ' مواجهة'] [Confrontation, global, Efficacy, Corona, Vaccine, Campaign] ['معبر', 'تدخل', 'الجيش', 'البوليساريو', 'مغاربة', "الكركرات"] [Guergarat, Moroccans, Polisario, Army, Intervention, crossing] ['الدمقراطي', 'الجمهوري', 'ترامب', 'الانتخابات', 'بايدن', 'الأمريكية'] [American, Biden, elections, Trump, Republican, Democrat]



Topic 0
Vaccination COVID
Topic 1
Events Guergart
Topic 2
USA elections
Topic 3
Football
Topic 4
drugs smuggling
Topic 5
Politics

Figure 6. Topics sampled by AraBERTopic.

Finally, we present in Figure 6 the wordcloud of the first six topics extracted by our model AraBERTopic.

5. CONCLUSION AND FUTURE WORK

We proposed in this paper a contextualized and neural AraBERT Topic Model (we named AraBERTopic) which integrates contextual knowledge to the neural topic model to capture more coherent and meaningful topics published in pages on the net.

For that, we collected 81 598 Arabic posts from Facebook pages of 7 Moroccan electronic press newspapers, then we preprocessed our dataset and extracted features using different embeddings models (Glove, Doc2Vec, and Asafaya – Arabic BERT). Finally, we extracted hidden topics in these pages using ProdLDA as a neural topic model and standard LDA.

To verify our contributions quantitatively, we performed experiments in terms of perplexity, topic coherence, and NPMI measures. The results proved that our proposed model can effectively capture meaningful topics and enhance the performance of topic modeling.

As part of our future work, we aim to enrich our ArabBERTopic model with new components to apply it to other tasks such as sentiment analysis using different deep learning algorithms (CNN, LSTM . . .).

References

[1] "Social Media users in Morocco - January 2021." [Online]. Available: https://napoleoncat.com/stats/social-media-users-in-morocco/2021/01

[2] S. K. Ray, A. Ahmad, and C. A. Kumar, "Review and Implementation of Topic Modeling in Hindi," *Applied Artificial Intelligence*, vol. 33, no. 11, pp. 979–1007, Sep. 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/08839514.2019.1661576

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] A. Srivastava and C. Sutton, "Autoencoding Variational Inference For Topic Models," *arXiv:1703.01488 [stat]*, Mar. 2017, arXiv: 1703.01488. [Online]. Available: http://arxiv.org/abs/1703.01488

[5] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2054–2059.

[6] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, Jun. 2019. [Online]. Available: http://link.springer.com/10.1007/s11042-018-6894-4

[7] M. Polatgil, "Analyzing comments made to the duolingo mobile application with topic modeling," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 223–230, Jan. 2023. [Online]. Available: https://doi.org/10.12785/ijcds/130118

[8] A. R. Peixoto, A. de Almeida, N. António, F. Batista, and R. Ribeiro, "Diachronic profile of startup companies through social media," *Social Network Analysis and Mining*, vol. 13, no. 1, Mar. 2023. [Online]. Available: https://doi.org/10.1007/s13278-023-01055-2

[9] M. R. Bhat, M. A. Kundroo, T. A. Tarray, and B. Agarwal, "Deep LDA : A new way to topic model," *Journal of Information and Optimization Sciences*, vol. 41, no. 3, pp. 823–834, Apr. 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/02522667.2019.1616911

[10] L. Liu, H. Huang, Y. Gao, X. Wei, and Y. Zhang, "Neural Variational Correlated Topic Modeling," p. 11.

[11] A. P. Logan, P. M. LaCasse, and B. J. Lunday, "Social network analysis of twitter interactions: a directed multilayer network approach," *Social Network Analysis and Mining*, vol. 13, no. 1, Apr. 2023. [Online]. Available: https://doi.org/10.1007/s13278-023-01063-2

[12] M. Makruf and A. Bramantoro, "Public hospital review on map service with part of speech tagging and biterm topic modeling," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 1097–1106, Apr. 2023. [Online]. Available: https://doi.org/10.12785/ijcds/130188

[13] X. Zhao, D. Wang, Z. Zhao, W. Liu, C. Lu, and F. Zhuang, "A neural topic model with word vectors and entity vectors for short texts," *Information Processing & Management*, vol. 58, no. 2, p. 102455, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S030645732030947X

[14] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse Additive Generative Models of Text," p. 8.

[15] D. Card, C. Tan, and N. A. Smith, "Neural Models for Documents with Metadata," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2031–2040, 2018, arXiv: 1705.09296. [Online]. Available: http://arxiv.org/abs/1705.09296

[16] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence," *arXiv:2004.03974 [cs]*, Apr. 2020, arXiv: 2004.03974. [Online]. Available: http://arxiv.org/abs/2004.03974

[17] Q. Xie, X. Zhang, Y. Ding, and M. Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *Journal of Informetrics*, vol. 14, no. 3, p. 101055, Aug. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1751157719305127

[18] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," pp. 9–15, May 2020. [Online]. Available: https://aclanthology.org/2020.osact-1.2

[19] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling," *arXiv:1312.3005 [cs]*, Mar. 2014, arXiv: 1312.3005. [Online]. Available: http://arxiv.org/abs/1312.3005

[20] M. Mancosu and F. Vegetti, "What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data," *Social Media + Society*, vol. 6, no. 3, p. 205630512094070, Jul. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2056305120940703

[21] MagedSaeed, "farasapy: A Python Wrapper for the well Farasa toolkit." [Online]. Available: https://github.com/MagedSaeed/farasapy

[22] "google-research/bert," Apr. 2021, original-date: 2018-10-25T22:57:34Z. [Online]. Available: https://github.com/google-research/bert

[23] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[24] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *arXiv:1405.4053 [cs]*, May 2014, arXiv: 1405.4053 version: 2. [Online]. Available: http://arxiv.org/abs/1405.4053

[25] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539.

Mrs. Nassera HABBAT PhD student in Hassan II University Morocco, obtained his Master degree in Information Systems Engineering from Caddi Ayyad University, and Licence degree in Technology and Web Programming from the same University in 2015 and 2013 respectively. Her research interest are: Big data and Machine learning.

Dr. Houda ANOUN received her engineering degree in Software Engineering from ENSEIRB Bordeaux in 2003 and her PhD in computational linguistics from Bordeaux I University in 2007. And actually she is a professor in the Computer Science Department at Ecole Supérieure de Technologie (Hassan II University), where she has been since 2009. Her current research interest lie in the area of IA especially deep learning, machine learning, and Big Data.

Dr. Larbi Hassouni got his engineer degree in 1983 from the "École Centrale de Marseille". He prepared his Phd degree in 1987 at the University of Aix Marseille III in France. Among his research work, there is the development of a list unification software with LeLisp language of INRIA in order to contribute to the realization of the inference engine of an expert system for the digital circuits' diagnossis. He also developed a behavioral symbolic simulator of digital circuits using the C, FRL, and LeLisp languages in order to contribute to the development of a "formal proof" tool for correcting a design of material produced by a Hardware Design Language (HDL). Currently, his research work mainly concerns Data Sciences.