EXTRACTING TOPICS FROM A TV CHANNEL'S FACEBOOK PAGE USING CONTEXTUALIZED DOCUMENT EMBEDDING

Nassera HABBAT[1*], Houda ANOUN[2], and Larbi HASSOUNI[3]

[1] RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco, nassera.habbat@gmail.com
[2] RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco, houda.anoun@gmail.com
[3] RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco, lhassouni@hotmail.com

KEY WORDS: AraBERT, ELMO, Neural topic model, LDA, ProdLDA, Topic coherence.

ABSTRACT:

Topic models extract meaningful words from text collection, allowing for a better understanding of data. However, the results are often not coherent enough, and thus harder to interpret. Adding more contextual knowledge to the model can enhance coherence. In recent years, neural network-based topic models become available, and the development level of the neural model has developed thanks to BERT-based representation. In this study, we suggest a model extract news on the Aljazeera Facebook page. Our approach combines the neural model (ProdLDA) and the Arabic Pre-training BERT transformer model (AraBERT). Therefore, the proposed model produces more expressive and consistent topics than ELMO using different topic model algorithms (ProdLDA and LDA) with 0.883 in topic coherence.

1. INTRODUCTION

Nowadays, because of the exponential development of the Internet, millions of data are generated every day, such as online news, in particular on social media like Facebook that is one of the largest used social media in the world. ("Social Media users," 2021.); Facebook had 2.85 billion active users each month at the end of 2020. Thus, Topic modeling is a way of text mining to discover hidden topics in social media.

Recently, neural topic models have become available with interesting results such as ProdLDA, which is an extended version of Latent Dirichlet Allocation (LDA) using deep learning, it replaces the mixture model in LDA with a product of experts to produce more coherent topics.

A good topic model is supposed to identify a group of meaningful words, and pre-trained word representations are a key component in many neural languages understanding models by adding contextual information to neural topic models to enrich the representations. Bidirectional Encoder Representations from Transformers (BERT) is the most outstanding of these models.

This paper's main contribution is to propose a model to pull out topics from Arabic information that is posted on the Aljazeera Facebook page using the Arabic pre-training BERT transformer model (AraBERT) as a word embedding model and ProdLDA as a neural topic model. we compared in text representation step; AraBERT with the contextual word embedding model: Embeddings from Language Models (ELMO), after we compared in topic modeling step ProdLDA with classic LDA. The results of our experiments show that the proposed model outperforms baseline models in terms of topic coherence, Normalized Pointwise Mutual Information (NPMI), and perplexity evaluation metrics.

This paper first presents a brief literature review in Section 2. The proposed model is illustrated in Section 3. In section 4, we present the results of our experiments. We end with a conclusion where we summarize this paper and outline our future work.

2. RELATED WORK

In this section, we exhibit some methods of topic extraction that include both traditional text representation methods and deep learning methods.

Topic modeling is one of the most difficult and promising tasks in natural language processing (NLP), and LDA is one of the most popular probabilistic topic models due to its nice generalization ability and extensibility. In (Jelodar et al., 2019), some authors studied scholarly articles published between 2003 and 2016 concerning topic modeling using LDA.

Recently, the Deep learning becomes a strong machine learning technique. It can learn multiple presentation layers and diverse processes exist especially adapted for learning expressive hidden representation. Amongst these methods; the Variational Autoencoder (VAE) which is a deep generating model. In this context, the authors expose in (Srivastava and Sutton, 2017) the first effective autoencoding variational Bayes (AEVB), it is a reasoning method using (LDA) to get the AVITM model, As an illustration of this, they replaced the mixture model in LDA with a product of experts to get a new topic model named ProdLDA. To evaluate their model they applied it in 20 Newsgroup dataset and they concluded that AVITM is superior to the traditional method in terms of accuracy, the reasoning time is longer, and ProdLDA provides more interpretable topics.

Only some recent researches have employed contextual embedding models like Bidirectional Encoder Representations from Transformers (BERT), Universal Language Model Fine-tuning (ULMFiT), and ELMO, which represent words and sentences, taking into consideration the contextual information within texts.

Using ULMFiT structure and Wikipedia resources, the authors in (ElJundi et al., 2019) developed a Universal Language Model

(hULMonA) which outperformed various state-of-art Arabic sentiment analysis datasets. On another hand, (Antoun et al., 2020.) presented the pre-trained AraBERT for the Arabic language (AraBERT), and they applied it to several natural language understanding tasks and compared it with multilingual BERT (Devlin et al., 2019) and it attained state-of-the-art performance in different NLP tasks concerning the Arabic language.

3. PROPOSED MODEL

In this section, we will describe the overall architecture of our model as shown in Fig.1 the model comprises three principal steps:

(1) Data collection: In this step, the web scraping techniques were used to pull posts from the Aljazeera Facebook page;

(2) AraBERT embedding: This step aims to pre-trained the model and extract the features;

(3) Topic Modeling: The last step consists of mining topics using Neural ProdLDA.

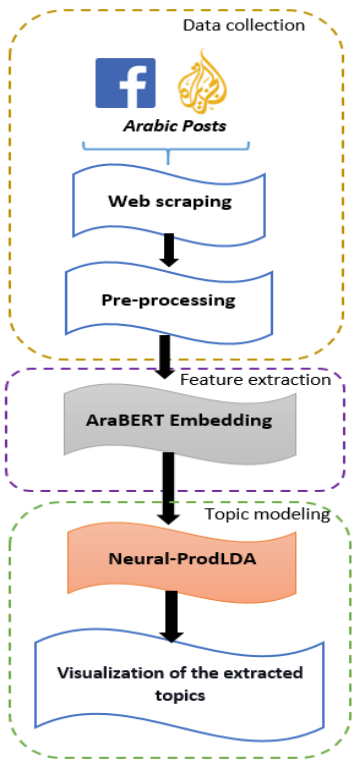

Figure 1. Flow Diagram of proposed approach.

3.1 Text representation

Word embedding is the collective name for a set of feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of numbers. In our study, we used BERT as word embedding model.
AraBERT (Antoun et al., 2020.) is an Arabic pre-training BERT transformer model. It makes use of a Transformer, which learns contextual relations between words in a text. The Transformer encoder is regarded as bidirectional because it reads the entire suite of words at once as opposed to directional models, which read the text input sequentially (right-to-left or left-to-right).

AraBERT was assessed on three downstream tasks of understanding the Arabic language: Sentiment Analysis, Named Entity Recognition, and Question Answering.
The sequence of tokens is the input, which is firstly embedded into vectors, and after that, the neural network processed the vectors to get the output as the sequence of vectors. where [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token.
The learning of AraBERT is divided into two stages as shown in figure 2:

1. The pre-training phase: it is only done once; it creates a neural network that has some general understanding of the used language.
2. The fine-tuning step: It enables the network to receive training on determined NLP tasks like question answering, classification, and Topic modeling.

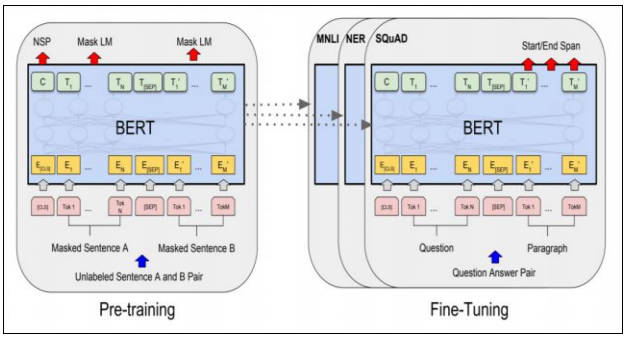

Figure 2. Process for BERT (Devlin et al., 2019) .

We used in our study the Arabic embedding ‘pretrained bert-base-arabert’ with 768 hidden dimensions, 12 encoder layers, 12 attention heads, and 110 M parameters, it is available from the Google BERT model site (*google-research/bert*, 2021)

3.2 Neural Topic Model

ProdLDA (Srivastava and Sutton, 2017) comes to solve a problem with the distribution $p(w|\theta,\beta)$ in LDA that is a mixing of multinomials, this is consists of don't make any forecasts that are sharper than the constituents that are being mixed, and this produces some topics with poor quality and does not correspond well with human judgment.
ProdLDA uses a Variational Autoencoder (VAE) on LDA employing a Laplace estimation function for the distribution of Dirichlet, thus allowing the training of a Dirichlet VAE. In addition, the ProdLDA model doesn't reparametrize the Dirichlet distribution directly.

Concerning ProdLDA and LDA parameters, we chose 15 clusters that represent topics and the top-ten topic words for each topic. As for the topics' number (K), we tried various K values from five to 80. Among these K values, the results were clearest when the K value was 15 with the highest topic coherence value.

4. EXPERIMENTS

We will describe in this section collect and preprocessing of our dataset, and then we will introduce baseline models and the used evaluation metrics.

4.1 Dataset

We used web scraping techniques with Version 3.8 of Python language to retrieve data from the Facebook page by using external libraries designed for automatic browsing specifically Version 2.25.0 of Requests and Version 4.9.3 of BeautifulSoup4.
We were interested in our study on Arabic information posted on the Aljazeera Facebook page. We collected 295,754 posts published from 14 August 2020 to 15 May 2021.
The collected posts were stored in the MongoDB database and pre-processed following these steps:

- Removal of Arabic stopwords, hyperlinks, hashes to keep only Arabic text.
- Lemmatization, it is the process of converting every word in a text to its base form using Farasapy Arabic Lemmatizer (MagedSaeed, n.d.).
- Creating the vocabulary comprising the most prevalent 2,000 terms.

4.2 Experimental setup

We used in our experiments, the implementation in Version 1.6.0 of Pytorch Library, we employed Adam as an optimizer with a learning rate of 2e-3 and batch size of 64. We fine-tuned our model after 20 epochs on the data. The experimental parameter settings are summarized in the following table 1:

Parameter	Value
Batch size	64
Number of Components	15
Number of epochs	20
Activation	softplus
Hidden Sizes	(100, 100)
Optimizer	Adam
Learning Rate	0.002
Dropout	0.2
Reduce On Plateau	False
momentum using for training	0.99

Table 1. Settings of the implemented model.

4.3 Baseline models

We compared our model with other baseline models to demonstrate the performance of the proposed model:
In word embedding level, we compared AraBERT model with ELMO as a contextual embedding model.
ELMO (Peters et al., 2018) is a deep contextualized word representation that simulates :

- The complexity of word use such as semantics and syntax.
- How do these usages change in different language environments (to model polysemy)

These word vectors are the internal state learning functions of the deep bi-directional language model (biLM) pre-trained on a large text corpus. They can be added to the existing models, which greatly improves the state of the art across a broad range of challenging NLP problems, including textual entailment, question answering, and sentiment analysis.
All ELMO models were trained on the 1 Billion Word Benchmark (Chelba et al., 2014) (~ 800M tokens of news crawl data from WMT 2011).
Concerning topic modeling level, we compared ProdLDA with LDA using ELMO and AraBERT word embedding models.

4.4 Metrics

The performance of the topic model is mainly evaluated as follows:

1. Subject consistency based on Normalized Pointwise Mutual Information (NPMI) algorithm,
2. topic coherence measure,
3. Perplexity evaluation.

4.4.1 NPMI: NPMI (Lau et al., 2014) calculates an automatic measurement of topic quality to assess our model with baselines. It comes from Pointwise Mutual Information (PMI), it measures how much one term w_i tells about the other w_j. It is formally defined as follows:

$$PMI(w_i,w_j) = \log \frac{p(w_i,w_j)}{p(w_i)p(w_j)} \quad (1)$$

$p(w_j)$: The probability that the word w_j appears in the corpus,
$p(w_i,w_j)$: The probability-that the word w_i appears together with the word w_j in the corpus.

We took in our experiment, top-five topic words and compute the NPMI score for each word set with the equation:

$$NPMI = \sum_{m=1}^{j} \sum_{n=m+1}^{j} \frac{PMI(x_m, x_n)}{-\log P(x_m, x_n)} \quad (2)$$

Topics with higher NPMI scores are the further probable terms to appear in the actual document m more frequently.

4.4.2 Topic coherence: Like NPMI, topic coherence measure also comes from PMI, it scores a topic by computing the degree of semantic similarity between its high scoring words. It is computed as follows:

$$Score_{CV}(w_m, w_n) = \log \frac{p(w_m, w_n) + \varepsilon}{p(w_m)p(w_n)} \quad (3)$$

4.4.3. Perplexity: Perplexity or confusion degree is a statistical measure to judge the quality of the subjects' modeling. It is defined as follows:

$$Perplexity(w|z,\theta,\beta) = \exp\left(\frac{-\sum_{d=1}^{D}\sum_{n=1}^{N_d}\log p(w_{dn}|z_{dn},\theta_d,\beta)}{\sum_{d=1}^{D}N_d}\right) \quad (4)$$

Where N_d is the number of terms in the document D.
θ is the density of document-topic.
β is the density of topic-word.

4.5 Results analysis

4.5.1. Quantitative Evaluation: To further compare the performance of various topic modelling algorithms with different word embedding models on short Arabic text data, we computed of each model; topic coherence, NPMI, and perplexity measures. As shown in Figure 3, the NPMI of the proposed model (AraBERT+ProdLDA) is higher than the other models with 0.853, Generally as shown in Table 2, our model is superior to the reference methods in terms of NPMI, topic coherence value, and perplexity score (the lowest value is the best).

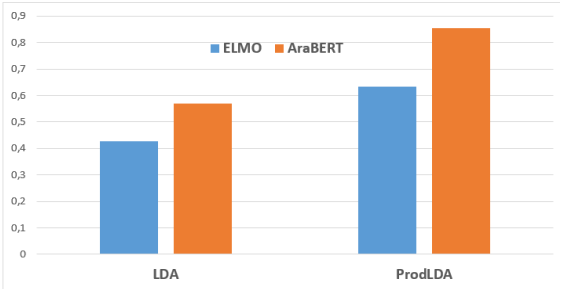

Figure 3. NPMI of different models.

Models		Metrics		
Topic models	**Word embeddings**	**NPMI**	**Topic coherence**	**Perplexity**
ProdLDA	***AraBERT***	**0.853**	**0.883**	**8.089**
	ELMO	0.633	0.710	9.436
LDA	***AraBERT***	0.570	0.696	11.25
	ELMO	0.428	0.594	20.68

Table 2. The comparison result of topic models with different embedding models on the collected dataset.

4.5.2. Qualitative Evaluation: To demonstrate the goodness of the topics produced by the proposed model, we present in Table 3, the top five words of three topics from all the models with the translation of each term in English.

The topics extracted by the proposed model are more coherent. In the second place, we see topics generated by ProdLDA + ELMO that seem to be coherent, but are affected by some added mixture of topics; for instance, the 1st topic is about COVID-19 but it contains the word 'March', and it is a little different from the subject.

Models	Topics
ProdLDA + AraBERT	['فلسطين', 'تصعيد', 'القدس', 'القوات', ' , ' إسرائيل'] ['Palestine', 'Escalation', 'Jerusalem', 'Forces', 'Israel'] ['مواجهة', 'الفيروس', 'كورونا' , ' الوقاية', 'لقاح'] [Confronting', 'virus', 'corona', 'prevention', 'vaccine'] ['لبنان', 'مرفأ', 'بيروت', 'انفجار', 'قتلى'] ['Lebanon', 'port', 'Beirut', 'explosion', 'dead']
ProdLDA + ELMO	['مسيرة', 'المواطنين', 'حملة', 'كورونا', 'الحجر'] ['march', 'Quarantine', 'Citizens', 'Campaign', 'Corona'] ['أفغانستان', 'طالبان', 'أمريكا', 'هجوم', 'محادثات'] ['Afghanistan', 'Taliban', 'America', 'attack', 'talks'] ['فلسطين', 'احتلال', 'المستوطنين', 'الاقصى', ' , ' إسرائيل'] ['Palestine', 'Occupation', 'Settlers', 'Al-Aqsa', 'Israel']
LDA + AraBERT	['كورونا', 'الصيني', 'بريطانيا', 'لقاح', 'الجغرافي'] ['Geographic', 'Vaccin', 'Britain', 'Chinese', 'Corona'] ['القوات', 'موريتانيا', 'الجزائر', 'الملكية', 'المسلحة"] ['Mauritania', 'forces', 'armed', 'royal', 'Algeria'] ['ترامب', 'النواب', 'بايدن', ' أمريكا ', 'مشروع'] ['Bill', 'America', 'Biden', 'Representatives', 'Trump']
LDA + ELMO	['السياسية', ' مواجهة', 'حملة', 'كورونا', 'فعالية'] ['Confrontation', 'global', 'Efficacy', 'Corona', 'Political'] ['الحريري', 'مرفأ', 'التجارية ', 'انفجار', 'انتخابات'] ['Hariri', 'Port', 'Commercial', 'Explosion', Elections'] ['الجمهوري', 'ترامب', 'الانتخابات', 'شخصية', 'الأمريكية'] ['American', 'Personality', 'elections', 'Trump', 'Republican']

Table 2. Top five words of three topics extracted by compared models.

5. CONCLUSION AND FUTURE WORK

In this work, we proposed a context-sensitive and neural AraBERT Topic Model, which enriches the model with more contextual knowledge to have more consistent and expressive topics.

In this context, we collected 295,754 posts in the Arabic language from the Aljazeera Facebook page, then this dataset was pre-processed, and to extract features, we used two contextual word embeddings models (AraBERT and ELMO). Finally, we used a neural topic model ProdLDA and classic LDA to generate hidden topics on this page.

To examine our model, we conducted experiments on NPMI, topic coherence, and perplexity evaluation metrics. The experimental outcomes show that the proposed approach (AraBERT + ProdLDA) is effective to improve the performance of topic modeling and capture meaningful topics.

As part of our future work, we plan to take advantage more of BERT by applying it to other NLP tasks, namely sentiment analysis tasks, and combining the BERT model with various deep learning algorithms (LSTM, Bi-LSTM, GRU...).

REFERENCES

Antoun, W., Baly, F., Hajj, H., n.d. AraBERT: Transformer-based Model for Arabic Language Understanding 7.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T., 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. ArXiv13123005 Cs.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs.

ElJundi, O., Antoun, W., El Droubi, N., Hajj, H., El-Hajj, W., Shaban, K., 2019. hULMonA: The Universal Language Model in Arabic, in: Proceedings of the Fourth Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Florence, Italy, pp. 68–77. https://doi.org/10.18653/v1/W19-4608

google-research/bert, 2021. . Google Research.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl. 78, 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Lau, J.H., Newman, D., Baldwin, T., 2014. 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. Presented at the EACL 2014, Association for Computational Linguistics, Gothenburg, Sweden, pp. 530–539. https://doi.org/10.3115/v1/E14-1056

MagedSaeed, n.d. farasapy: A Python Wrapper for the well Farasa toolkit.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. ArXiv180205365 Cs.

Social Media users - January 2021 [WWW Document], n.d. URL https://napoleoncat.com/stats/(accessed 2.18.21).

Srivastava, A., Sutton, C., 2017. Autoencoding Variational Inference For Topic Models. ArXiv170301488 Stat.