Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 189 (2021) 191–194

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique

Abeer Abuzayed and Hend Al-Khalifa*

iWAN Research Group, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Abstract

Topic modeling is an unsupervised machine learning technique for finding abstract topics in a large collection of documents. It helps in organizing, understanding and summarizing large collections of textual information and discovering the latent topics that vary among documents in a given corpus. Latent Dirichlet allocation (LDA) and Non-Negative Matrix Factorization (NMF) are two of the most popular topic modeling techniques. LDA uses a probabilistic approach whereas NMF uses matrix factorization approach, however, new techniques that are based on BERT for topic modeling do exist. In this paper, we aim to experiment with BERTopic using different Pre-Trained Arabic Language Models as embeddings, and compare its results against LDA and NMF techniques. We used Normalized Pointwise Mutual Information (NPMI) measure to evaluate the results of topic modeling techniques. The overall results generated by BERTopic showed better results compared to NMF and LDA.

Keywords: Topic modeling; BERT; BERTopic; LDA; NMF; NPMI; Arabic Language.

1. Introduction

Topic modeling is a task that clusters documents and words with similar meanings. It has important applications in many fields like Natural language processing (NLP) and Information Retrieval (IR). It uses unsupervised machine learning techniques to discover topics from document collections. Different topic modeling approaches are available starting with Probabilistic latent semantic analysis (PLSA) that was first proposed in 1999 [1], then the Latent Dirichlet Allocation (LDA) that was proposed in 2003[2], which then became one of the most popular approaches. Similarly, Non-negative matrix factorization (NMF) is an unsupervised approach for reducing the dimensionality of nonnegative

* Corresponding author: hendk@ksu.edu.sa

10.1016/j.procs.2021.05.096

matrices [3], and it has been widely used to discover the underlying relationships between texts and identify latent topics [4]. Such topic modeling approaches require no labels to operate, also they need to specify beforehand the number of categories to cluster around. Nonetheless, there is a growing number of topic modeling approaches that are based on LDA and NMF as a starting point, yet, they take quite some efforts through hyperparameter tuning to create meaningful topics.

On the other hand, advancement in Pre-trained Language Models (PLMs) have played a role in topic modeling task. For example, BERTopic [5] is a topic modeling technique that leverages BERT embeddings and a class-based TF-IDF to create dense clusters, it also uses Uniform Manifold Approximation and Projection (UMAP) technique to lower the dimensionality of the embeddings before clustering the documents. Initial experiments with BERTopic technique have shown promising results [6], therefore, in this paper, we aim to experiment with BERTopic technique using different PLMs and compare their results against well-known techniques such as LDA and NMF.

The rest of the paper is organized as follows: Section 2 presents previous work on topic modeling for Arabic language. Section 3 presents the method and dataset used in our experiments. Section 4 discusses the results of the experiments, and section 5 concludes the paper with final remarks.

2. Previous Work

Although Arabic topic modeling research has been evolving over the past years, its application in practice is still limited. In this section, we present a few recent works related to the area of topic modeling for Arabic language.
Rafea and GabAllah [7] provided a survey of different methods of topic detection and modeling techniques related to the Arabic domain. They found that most previous work used LDA for topic modeling, however, recent work started to combine LDA with other techniques such as K-means clustering and word2vec embeddings. Recent work by [8] used paragraph vectors to create fixed-length vector representations for each verse/sentence in the Quran. They evaluated the derived clusters of related verses against a tagged corpus to verify the relationships between the verses of the Quran, identify how they are related, and address the concepts covered in each cluster. In the domain of social media of Arabic Web, topic modeling research has also focused on social media content, AlShalan et al. 2020 [9] used NMF to discover the main issues and topics discussed in hate tweets during COVID-19 pandemic. NMF helped them identify seven commonly discussed topics in hate tweets during the pandemic.

3. Method

In this study, we conducted a set of experimentations with different topic modeling techniques such as LDA, NMF and BERTopic. We started with LDA and NMF as baselines and then use BERTopic with various word embedding representations with topics numbers starts from 5 to 500 topics. We used Sklearn's implementation of NMF, Gensim implementation of LDA Multicore, and Google Colaboratory Pro for running the experiments. Our implementations are freely available online on Github for the research community[1].

3.1. Dataset

We used "DataSet for Arabic Classification" [10] which contains 111,728 Arabic documents written in Modern Standard Arabic (MSA). The dataset was collected from three Arabic online newspapers: Assabah, Hespress and Akhbarona. The documents in the dataset are categorized into 5 classes: sport, politics, culture, economy and diverse. We removed 2939 missing documents and ran the experiments with the remaining 108789 documents without any document labels. The text contains only alphabetic, numeric and symbolic words, so we have not applied any preprocessing as the text is almost clean.

1 https://github.com/iwan-rg/Arabic-Topic-Modeling

3.2. Evaluation metrics

Evaluating a topic model is arguably a challenging task given the nature of unsupervised models and the absence of standard measures and well-established tools. We evaluated our model topics coherence using Normalized Pointwise Mutual Information (NPMI) [11] where it measures the topic coherence between high scoring words in the topic. NPMI ranges from [-1,1] and it measures how much the top-10 words of a topic are related to each other, where higher positive NMPI is better. The NPMI was implemented using the Gensim library.

3.3. Baseline

We compared the results of BERTopic with two of the most popular topic modeling techniques: LDA and NMF. As these two models require providing the number of topics in advance, we decided to start our experiment with a minimum number of topics equal to the dataset classes number. Thus, we started with 5 topics, then increased the number gradually and reported the NPMI for each experiment. We stopped experimenting at 500 topics due to computational power limitations.

3.4. BERTopic Model

Mainly, our experiments are based on the open-source BERTopic model and experimenting with various word embedding representations such as pre-trained language models and AraVec2.0 [12] word embedding model. These embedding models are used to extract the contextualized word representation for all tokens and then passes it to BERTopic. BERTopic extracts the document embedding then uses both the UMAP algorithm to reduce the embedding dimensionality and the density-based algorithm HDBSAN for documents clustering [5]. In addition, it provides various options to extract document embeddings. Firstly, we can use sentence-transformers package [13] with two default models: "distilbert-base-nli-stsb-mean-tokens" for the English language and "xlm-r-bert-base-nli-stsb-mean-tokens" for any language other than English, where XLM-R models support 50+ languages. Secondly, the Flair framework [14] can be used to extract the document embedding for any Hugging Face transformers model or almost any embedding model available publicly. For the purpose of this study, we used the following pre-trained language models: AraBERTV2.0 [15], ARBERT [16], QARiB [17], XLM-R [18] and AraVec2.0. AraBERTV2.0, ARBERT and QARiB are monolingual Arabic BERT models, while XLM-R is multilingual. Furthermore, BERTopic does not require defining the number of topics in advance where it can extract the number of topics described in the documents. Although this is an advantage of BERTopic over LDA and NMF, however, for the purpose of this study we ran the models with a similar number of topics ranging from 5 to 500 topics.

4. Results and Discussion

As we mentioned before we started our experiments with two baselines, namely LDA and NMF. As shown in Figure 1 (a), we reported NPMI value for LDA on two document sizes ~100k and 50k, this is because when we ran the experiment on the ~100k document size the LDA method stopped working at 350 topics due to the limitation of the computational power we used, on the other hand when we ran the same experiment on the ~50K (half the size of the documents dataset) we were able to reach 500 topics. Notice that on both experiments, LDA produced negative NPMI values which indicate the low performance of the model.

On the other hand, Figure 1 (b) illustrates the performance of the rest of the models. AraBERT, ARBERT and XLM-R preserve almost similar behaviour after increasing the number of topics to more than 150, however, AraVec outperformed them all. While NMF and QARiB behave completely differently. This is likely because of the difference in nature of the datasets each model has been trained on. AraBERT, ARBERT and XLM-R have been trained using a collection of MSA datasets which are mostly from Wikipedia, news, books and OSCAR corpus. In contrast, QARiB has been trained on a collection of Arabic tweets and sentences written in MSA. Similarly, AraVec2.0 has been trained on tweets, web pages and Wikipedia articles. Moreover, the models best performances are captured between topics number 100 to 200. Finally, the variation of models performance starting with 5 topics, which represents the dataset

original categories size, shows the probable disagreement between human judgements and machines in text categorization.

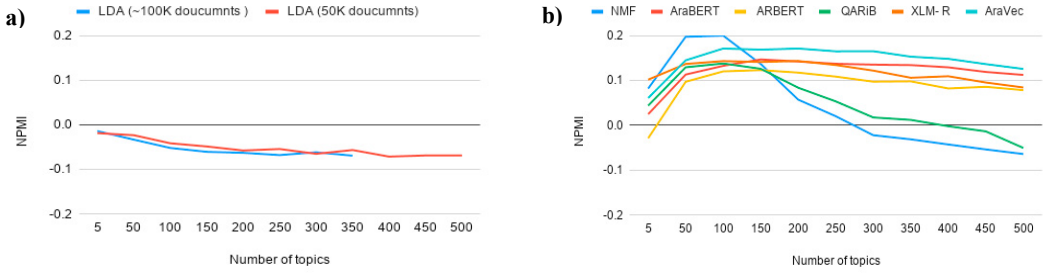

Fig. 1. (a) NPMI for LDA with different dataset sizes; (b) NPMI for NMF and BERTopic with various word embeddings.

5. Conclusion

In this paper, we presented a pilot study that experimented with BERTopic using different pre-trained Arabic language models as embeddings, and compared its results against LDA and NMF techniques. Despite the initial promising results of BERTopic and its flexibility of adopting to any topic size, yet finding an accurate measure for evaluating the quality of the generated topics is still challenging and it needs further research.

References

[1] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Mach. Learn., vol. 42, no. 1, pp. 177–196, Jan. 2001, doi: 10.1023/A:1007617005950.
[2] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, no. Jan, pp. 993–1022, 2003.
[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.
[4] S. Arora, R. Ge, and A. Moitra, "Learning Topic Models – Going beyond SVD," in 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, Oct. 2012, pp. 1–10, doi: 10.1109/FOCS.2012.49.
[5] Maarten Grootendorst, BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. Zenodo, 2021.
[6] M. Grootendorst, "Topic Modeling with BERT," Medium, Oct. 06, 2020. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6 .
[7] A. Rafea and N. GabAllah, "Topic Detection Approaches in Identifying Topics and Events from Arabic Corpora," Procedia Comput. Sci., vol. 142, pp. 270–277, Jan. 2018, doi: 10.1016/j.procs.2018.10.492.
[8] M. Alshammeri, E. Atwell, and M. A. Alsalka, "Quranic Topic Modelling Using Paragraph Vectors," in Intelligent Systems and Applications, Cham, 2021, pp. 218–230, doi: 10.1007/978-3-030-55187-2_19.
[9] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, "Detection of Hate Speech in COVID-19–Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach," J. Med. Internet Res., vol. 22, no. 12, p. e22609, Dec. 2020, doi: 10.2196/22609.
[10] M. Biniz, "DataSet for Arabic Classification," vol. 1, Mar. 2018, doi: 10.17632/v524p5dhpj.1.
[11] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction", 2009. https://www.semanticscholar.org/paper/Normalized-(pointwise)-mutual-information-in-Bouma/15218d9c029cbb903ae7c729b2c644c24994c201.
[12] A. B. Soliman, K. Eissa, and El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," Procedia Comput. Sci., vol. 117, pp. 256–265, Jan. 2017, doi: 10.1016/j.procs.2017.10.117.
[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," ArXiv190810084 Cs, Aug. 2019, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/1908.10084.
[14] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, Minnesota, Jun. 2019, pp. 54–59, doi: 10.18653/v1/N19-4010.
[15] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, May 2020, pp. 9–15, Accessed: Mar. 12, 2021. [Online]. Available: https://www.aclweb.org/anthology/2020.osact-1.2.
[16] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," ArXiv210101785 Cs, Dec. 2020, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/2101.01785.
[17] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations," ArXiv210210684 Cs, Feb. 2021, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/2102.10684.
[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," ArXiv190810084 Cs, Aug. 2019, Accessed: Mar. 12, 2021. [Online]. Available: http://arxiv.org/abs/1908.10084.