Arabic Topic Classification in the Generative and AutoML Era

Doha Albared and **Hadi Hamoud** and **Fadi A. Zaraket**
Arab Center for Research and Policy Studies, Doha
{dal007,hhamoud,fzaraket} @dohainstitute.edu.qa

Abstract

Most recent models for Arabic topic classification leveraged fine-tuning existing pre-trained transformer models and targeted a limited number of categories. More recently, advances in automated ML and generative models introduced novel potentials for the task. While these approaches work for English, it is a question of whether they perform well for low-resourced languages; Arabic in particular. This paper presents (i) ArBNTopic; a novel Arabic dataset with an extended 14-topic class set covering modern books from social sciences and humanities along with newspaper articles, and (ii) a set of topic classifiers built from it. We fine-tuned an open LLM model to build ArGTC. We compared its performance against the best models built with Vertex AI (Google), AutoML(H2O), and AutoTrain(HuggingFace). ArGTC outperformed the VertexAi and AutoML models, and was reasonably similar to the AutoTrain model.

1 Introduction

Text classification models have been a topic of interest in the natural language processing (NLP) research community, due to their importance in performing multiple tasks such as sentiment analysis (Sohangir et al., 2018; Qian et al., 2018; Ain et al., 2017), topic classification (Johnson and Zhang, 2017; Razno, 2019), spam detection (Trivedi, 2016; Ismail et al., 2022; Fattahi and Mejri, 2021), and fake news detection (Meesad, 2021; Hamid et al., 2020; Kong et al., 2020). In earlier stages, text classification primarily relied on traditional machine-learning algorithms like Support Vector Machines (SVM), Naive Bayes, and Decision Trees. Deep Learning models have been used to perform numerous NLP tasks and attained remarkable results (Sohangir et al., 2018; Lai et al., 2015). Nonetheless, these models are built from the ground up, demanding large datasets and several days of training.

Recently, transfer learning by fine-tuning a pre-trained large language model (LLM) helped solve the problem. An LLM, trained with large corpora for generic tasks, gains further specific capacities in the process. This effectively produced high-performing classification models across several languages (Adhikari et al., 2019; Balkus and Yan, 2022b; Bataa and Wu, 2019; Polignano et al., 2019).

Recently, state-of-the-art models tailored for Arabic NLP tasks leveraged transfer learning to perform tasks such as text generation, classification, translation, sentiment analysis, summarization, title generation, and dialect identification. AraBERT, one of the most prominent models (Antoun et al., 2020a), was fine-tuned from BERT (Devlin et al., 2018) to specifically serve the Arabic language. More models for Arabic emerged to perform text generation and classification (Nagoudi et al., 2021; Khondaker et al., 2023; Khered et al., 2022), and topic classification: identifying the topic(s) discussed in a specific text (Abdul-Mageed et al., 2020; Chowdhury et al., 2020).

Generative Models: More recently, several generative transformer-based models have been trained on multi-lingual corpora including Arabic. We considered two fine-tuned variants: BLOOMZ and mT0 (Muennighoff et al., 2022b). Bloomz-7b-mt (Muennighoff et al., 2022b) is a 7-billion parameter model pre-trained to respond to instructions (z) in further languages leveraging the xPmt multilingual (mt) dataset (Muennighoff et al., 2022a) with significant Arabic content .

AutoML: Meanwhile, several automated machine learning services emerged such as Google Vertex Ai (Google-Vertex-AI, 2023),

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 399–404

H2O AutoML (LeDell and Poirier, 2020), and Huggingface AutoTrain (Wolf et al., 2020). Such services perform tasks including data preprocessing, feature engineering, model selection, hyperparameter tuning, and model deployment saving time, effort, and resources. These models perform typically well in text classification for high-resource languages.

ArBNTopic & ArGTC: In this paper, we explore Arabic text classification, with an extended set of topics, across the generative and automated machine learning approaches. For that we built ArBNTopic, a dataset from specialized books and newspaper articles to introduce novel topics to Arabic topic classification models including topics from sciences, social sciences, and humanities. This dataset is openly available on HuggingFace [1]

We used ArBNTopic to build ArGTC[2] in two steps. The first step boosted the bloomz-7b-mt model with additional Arabic content from domains it did not cover before. In the second step, we fine-tuned the resulting model [3] from step 1, with a part of ArBNTopic with classes. We chose the bloomz-7b-mt after a careful review as its predecessors had Arabic capacities (BLOOM), had instruction (Z) fine-tuning capacities, and had additional multilingual (mt) capacities from additional diverse datasets.

ArGTC performs with an accuracy, precision, and recall of 83, 81, and 81%, respectively. It shows better results than the best models we generated with ArBNTopic using Google Vertex Ai (Google-Vertex-AI, 2023) and H2O AutoML (LeDell and Poirier, 2020). It also compares closely to the performance of the best model generated using AutoTrain from Huggingface (Wolf et al., 2020). ArGTC is reasonably better than the models generated using the automated machine learning services when considering a cost-effective performance balance.

2 Related Work

The use of transformer-based models is quickly covering all NLP tasks. It started with BERT-architecture models (Li et al., 2022). Transformer models take advantage of their abilities to represent contextual relationships between concepts through contextual embeddings. Arabic text categorization research also emerged recently (Alammary, 2022). Several Arabic NLP tasks followed after the inception of multilingual BERT. However, studies reported better results using monolingual models, specifically for low to mid-resource languages (Wu and Dredze, 2020). AraBERT, trained on Arabic Wikipedia and newspaper, was applied to several downstream tasks (Antoun et al., 2020b). The model performed well on multi-class tasks (zahra El-Alami et al., 2022).

Training on a mixture of Dialect-Arabic through social media datasets, and modern standard Arabic (MSA) data, has resulted in better encoder models, like Qarib (Abdelali et al., 2021). Text categorization attempts on iterations of Qarib have proven successful, through fine-tuning on classified MSA and Dialect datasets, with 6 to 12 labeling classes used (Chowdhury et al., 2020).

Apart from BERT models, considerable progress has been made using GPT and T5-based models. For instance, AraT5, a fine-tuned version of multilingual T5 on the Arabic Language (Nagoudi et al., 2022), is achieving close to state-of-the-art performances on a variety of tasks, including categorization (Khondaker et al., 2023).

GPT-3.5/4 base models augmented with Arabic data, are also performing exceptionally well on text and sequence classification(Abdelali et al., 2023; Balkus and Yan, 2022a). However, due to OpenAi's business model, fine-tuned versions of GPT are only available for use through the paid API.

Also important to mention that with the release of the BLOOM family, including the BLOOMZ model, which is our choice of foundation, BigScience has also deployed multiple inference heads on top, with specific configurations, for different tasks. This includes sequence classification, question answering, and generation.

3 Fine-tuning Data and Model

We selected the 7-Billion parameter, publicly available, bloomz-7b-mt (Muennighoff et al., 2022b) model as our base model. It belongs to the BLOOMZ and mT0 family resulting from fine-tuning BLOOM on the multilingual

[1] https://huggingface.co/datasets/dru-ac/ArBNTopic
[2] https://huggingface.co/dru-ac/ArGTC
[3] https://huggingface.co/dru-ac/FTArBloom

xP3mt dataset (Muennighoff et al., 2022a). We fine-tuned it in two phases: (i) for Arabic text generation, and then (ii) for Arabic text classification.

To realize this objective we fine-tuned bloomz-7b-mt on an additional 58,682 tokens taken from books and newspapers datasets written in modern standard Arabic. The resultant model encompasses a comprehensive spectrum of fourteen distinct subjects that are Religion, Finance and Economics, Politics, Medical, Culture, Sports, Science and Technology, Anthropology and Sociology, Art and Literature, Education, History, Language and Linguistics, Law, as well as Philosophy.

The generated model was further fine-tuned on a bigger and labeled dataset that comprises 833,642 tokens. We call the resulting model ArGTC. ArGTC is designed to categorize input text, determining its alignment with one of the 14 predefined topics.

3.1 Data and Preprocessing

We developed ArBNTopic to fine-tune and train the generative and AutoML-based models. We used newspaper articles and a set of published books to build ArBNTopic. The newspaper articles come from the SANAD dataset (Einea et al., 2019). SANAD is publicly available and includes an extensive assortment of Arabic news articles suitable for various Arabic NLP tasks. These articles were gathered from three well-known Arabic news portals: AlKhaleej, AlArabiya, and Akhbarona. Each newspaper dataset is labeled with seven categories: Culture, Finance, Medical, Politics, Religion, Sports, and Tech, except for AlArabiya, which lacks the Religion category. We split the articles down into the paragraph level with a maximum character limit of 250 per segment. When a paragraph contained more than that, we split it into more than one segment and included all segments.

The dataset of books was acquired in Word format provided by the Arab Center for Research and Policy Studies. The books spanned the areas of Religion, Economy, Politics, Anthropology and Sociology, Art and Literature, Education, History, Language and Linguistics, Philosophy, and Law. Table 1 shows the number of books across each category. We selected texts from the books to build a balanced dataset across categories.

Table 1: Number of books by category

Category	Number of Books
Religion	8
Economy	28
Language & Linguistics	25
Anthropology & Sociology	101
Art & Literature	7
Education	6
Philosophy	84
Law	6
Politics	174
History	79

Newspaper articles provide concise and short texts with ideas and sentences that tend to be more compressed and more straightforward than texts in books. Texts in books tend to be quite the opposite; longer texts that feel free to tackle topics from broader and different angles. Hence, a book classified in one category can easily overlap in some of its parts or chapters with other categories. That is a book in history can discuss religion, politics, and education. A book in philosophy can discuss religion and society. When a sentence from a book is taken out of its context, it can be categorized into a topic other than the topic of its book.

At times, sentences from books may become too general to be categorized with any topic at all if read out of context. We resolved these issues by confining our data to selecting the first sentence(s) after each title, subtitle, and heading in each of the books. The first sentences after the titles tend to contain the thesis statements and topic sentences which introduces and summarizes the discussions in next paragraphs to follow.

We still had to solve the disparity in the number of books under each category. The first sentence after a title rule yielded a wide variation gap in the number of documents for each category. For example, while around 200 sentences were labeled religion, more than 4000 were labeled politics. To overcome this issue, for the categories that had less than 10 books (religion, law, art and literature, education), we extracted the first sentences of each long paragraph.

We unified the labels for categories from newspapers and books as shown in Table 2 to obtain the aggregated list of categories. All the data extracted from books is openly available on Hugginface. The sample mixed dataset

(books and newspapers) used to complete the finetuning is available as available as well.

Table 2: Categories mapping

Newspapers	Books	Final
Religion	Religion	Religion
Finance	Economy	Finance-and-Economy
Politics	Politics	Politics
Medical		Medical
Culture		Culture
Sports		Sports
Tech		Science-and-Technology
	Anthropology & Sociology	Anthropology-and-Sociology
	Education	Education
	Philosophy	Philosophy
	Language and Linguistics	Language-and-Linguistics
	History	History
	Law	Law

3.2 Compute Setup

We fine-tuned both models on one A100-80 GiB GPU that we rented from a provider on the cloud. The GPU had 80 GiB VRAM, 125 GiB RAM, 14 vCPUs, 256 GiB of persistent volume disk storage, and 256 GiB of container non-persistent disk storage. Compute power and storage costs of this configuration amount to $1.86/hour. The model can be loaded for inference on a system with a GPU of 32 GiB VRAM at a cost of $0.3 per hour.

We loaded and fine-tuned the 7-Billion parameter model with limited virtual memory due with the help of gradient checkpointing [4] [5]. This comes at the expense of slowing down the fine-tuning process.

As for software specifications, we used a standard deep-learning container image, with Python-3.10, PyTorch-2.0.1, Cuda-11.8.0, and Transformers-4.32.1.

4 Results

Following the fine-tuning and upon testing the ArGTC on ArBNTopic labeled with the 14 categories, it scored 83% accuracy, 81% precision, and recall (Table 3). In addition, we trained 3 more classification models using automated ML services from Google Vertex Ai, Hugginface AutoTrain, and H2O AutoML. In fact, Multiple H2O models were trained using embeddings as features from different language models trained on Arabic data, namely, AraBERT (Antoun et al., 2020b), and AraGPT2 (Antoun et al., 2021). Using AraGPT2, embeddings yielded a model with higher scores, 76% accuracy and precision, and 74% recall, then the models were trained using AraBERT embeddings, which scored 67% for accuracy, precision, and recall.

While the best H2O model scored lower than ArGTC on all three measurements, the model produced using AutoTrain scored 84% for accuracy, precision, and recall. In addition, the fourth classification model we trained using Google Vertex AI scored 77% precision, 76% recall, and 76% accuracy. Hence, the AutoTrain model scored the highest on all measurements, accuracy, precision, and recall, followed closely by ArGTC, then the H2O model followed by the model obtained through Google Vertex AI training.

It is worth noting that while H2O is charge-free, AutoTrain is free only for datasets up to 3,000 samples. For which contained 19,784 samples, extra charges amount to $46. On the other hand, using Vertex Ai with the same dataset costs $22. Fine-tuning time for ArGTC was 4-5 hours, which amounts to a cost of around $9.3. Consequently, the ArGTC model optimally balances performance against costs.

Table 3: Results of training ArGClass, Google Vertex Ai (Vertex), HuggingFace AutoTrain (HF), H2O AutoML (H2O)

	ArGTC	Vertex	H2O	HF
Accuracy %	83	76	76	84
Precision %	81	77	76	84
Recall %	81	76	74	84

5 Conclusion

This paper introduced ArGTC, an Arabic text classification model with 14 categories. Utilizing transfer learning techniques, the model was fine-tuned in two stages from bloomz-7b-mt, achieving 83% accuracy, 81% precision, and recall and surpassing the best models trained using VertexAi and AutoML. It performed comparably to the best model we trained with AutoTrain with a small margin.

[4] https://huggingface.co/docs/transformers/v4.18.0/en/performance#gradient-checkpointing

[5] https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9

Limitations

ArGTC performance is bound to the quality of ArBNTopic and the foundation models. ArBNTopic contains a specific collection of books. We also confined the models to a fixed list of predefined topics based on the specialties of the books and newspaper articles. To capture topics beyond this predefined set, alternative unsupervised topic modeling techniques would be necessary.

Ethics Statement

The data was collected and used with the appropriate approvals of the intellectual property owners. All results are reported following best academic standards and practices.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pretraining bert on arabic tweets: Practical considerations.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2023. Benchmarking arabic ai with large language models.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A Rehman. 2017. Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, 8(6).

Ali Saleh Alammary. 2022. Bert models for arabic text classification: A systematic review. *Applied Sciences*, 12(11).

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. 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*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Salvador Balkus and Donghui Yan. 2022a. Improving short text classification with augmented data using gpt-3.

Salvador V Balkus and Donghui Yan. 2022b. Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, pages 1–30.

Enkhbold Bataa and Joshua Wu. 2019. An investigation of transfer learning-based sentiment analysis in japanese. *arXiv preprint arXiv:1905.09642*.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J Jansen. 2020. Improving arabic text categorization using transformer training diversification. In *Proceedings of the fifth arabic natural language processing workshop*, pages 226–236.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Jaouhar Fattahi and Mohamed Mejri. 2021. Spaml: a bimodal ensemble learning spam detector based on nlp techniques. In *2021 IEEE 5th international conference on cryptography, security and privacy (CSP)*, pages 107–112. IEEE.

Google-Vertex-AI. 2023. Google vertex ai.

Abdullah Hamid, Nasrullah Shiekh, Naina Said, Kashif Ahmad, Asma Gul, Laiq Hassan, and Ala Al-Fuqaha. 2020. Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case. *arXiv preprint arXiv:2012.07517*.

Safaa SI Ismail, Romany F Mansour, Abd El-Aziz, M Rasha, Ahmed I Taloba, et al. 2022. Efficient e-mail spam detection strategy using genetic decision tree processing with nlp features. *Computational Intelligence and Neuroscience*, 2022.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Abdullah Salem Khered, Ingy Yasser Hassan Abdou Abdelhalim, and Riza Theresa Batista-Navarro. 2022. Building an ensemble of transformer models for arabic dialect classification and sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 479–484.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp.

Sheng How Kong, Li Mei Tan, Keng Hoon Gan, and Nur Hana Samsudin. 2020. Fake news detection using deep learning. In *2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE)*, pages 102–107. IEEE.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Erin LeDell and Sebastien Poirier. 2020. H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Phayung Meesad. 2021. Thai fake news detection based on information retrieval, natural language processing and machine learning. *SN Computer Science*, 2(6):425.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022a. Crosslingual generalization through multitask finetuning.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022b. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Arat5: Text-to-text transformers for arabic language generation. *arXiv preprint arXiv:2109.12068*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, et al. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CEUR Workshop Proceedings*, volume 2481, pages 1–6. CEUR.

Jun Qian, Zhendong Niu, and Chongyang Shi. 2018. Sentiment analysis model on weather related tweets with deep neural network. In *Proceedings of the 2018 10th international conference on machine learning and computing*, pages 31–35.

Maria Razno. 2019. Machine learning text classification model with nlp approach. *Computational Linguistics and Intelligent Systems*, 2:71–73.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25.

Shrawan Kumar Trivedi. 2016. A study of machine learning classifiers for spam detection. In *2016 4th international symposium on computational and business intelligence (ISCBI)*, pages 176–180. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert?

Fatima zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. Contextual semantic embeddings based on fine-tuned arabert model for arabic text multi-class categorization. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part A):8422–8428.