Mela at CheckThat! 2024: Transferring Persuasion Detection from English to Arabic - A Multilingual BERT Approach

Notebook for the CheckThat! Lab at CLEF 2024

Sara Nabhani, Md Abdur Razzaq Riyadh

Department of Artificial Intelligence, University of Malta

Abstract

This paper presents our system's participation in CheckThat! Lab Task 3, focuses on identifying persuasion techniques in Arabic text. We solely focused on Arabic, a low-resource language for this task. The task required identifying any persuasion technique applied to individual tokens within the text. Only the test set was provided for Arabic for this task, without any corresponding development or training sets. Our research aimed to investigate how a resource-rich language like English could benefit the low-resource Arabic language in the context of persuasion detection. To that end, we utilized a multilingual BERT which incorporated English and Arabic knowledge during its pre-training stage. Our system achieved first place on the Arabic leaderboard in the shared task. The result, achieved without training on Arabic data, highlights the effectiveness of multilingual BERT models. This also demonstrates the potential of using resource-rich languages like English to enhance performance in low-resource languages such as Arabic for persuasion detection tasks.

Keywords

arabic, propaganda, persuasion

1. Introduction

Throughout history, propaganda has played a significant role in shaping public opinion. Propaganda uses various persuasive techniques to influence the way people think and act. With the advent of the digital age, the impact of propaganda has grown even stronger. Nowadays, persuasive techniques are widely used as tools for spreading propaganda through digital platforms. The increasing use of these persuasion techniques highlights the need for advanced methods to identify and critically evaluate them. This need has become urgent as the volume of digital content continues to rise, making it easier for propaganda to spread rapidly.

This paper describes our approach to the CheckThat! task 3 [1, 2], focuses on the identification of persuasion techniques within the textual spans of Arabic articles. The goal of this task is to detect various techniques used to persuade readers within Arabic texts. However, a significant challenge we faced was the lack of training data for Arabic. While the task provided training data for several languages, including English, French, Italian, German, Russian, and Polish, there was no training set available for Arabic. This lack of training data for Arabic made it difficult to develop a model specifically trained on Arabic texts. To overcome this challenge, we used the training data from the English set to fine-tune a multilingual BERT model [3] and then evaluated it on the Arabic test set. Thus, our study investigates the effectiveness of using a high-resource language, such as English, to enhance the performance of a model for a low-resource language like Arabic. In the context of the persuasion technique identification task, we aimed to demonstrate that a model trained on English data could still perform effectively when applied to Arabic texts. This approach is based on the idea of cross-lingual transfer learning, where knowledge gained from one language can be transferred to another language.

The paper is structured as follows: Section 2 reviews previous works in this area. Section 3 outlines our proposed system in detail. Section 4 presents the results. Section 4 discusses our findings and their

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

sara.nabhani.23@um.edu.mt (S. Nabhani); md.riyadh.23@um.edu.mt (M. A. R. Riyadh)

implications. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Related Work

Persuasion detection has traditionally focused on analyzing entire documents or paragraphs. However, a fairly recent study introduced the task of identifying persuasion techniques at the token level [4]. Their work is significant because it provides one of the earlier datasets annotated with propaganda techniques at the character level. This allows researchers to employ multi-label, multi-class classification techniques for persuasion detection with finer granularity [4]. The author utilized BERT [3] for this downstream task and evaluated using a modified F1 score to consider partial matching.

Several recent studies have explored persuasion detection via shared tasks like SemEval and ArAIEval [5, 6, 7]. BERT-based classifiers are a popular choice for these tasks due to their effectiveness in text classification [5]. There is a challenge with label distribution, as some persuasion techniques appear much less frequently than others. Moreover, most tokens within the data lack any persuasion labels. This is addressed by employing techniques like class weighting during loss calculation [8]. Additionally, multi-task architectures utilizing shared representations from pre-trained models like BERT have shown good results [9]. For persuasion detection in Arabic, previous works are commonly based on AraBERT [10] [7]. Propaganda detection in Arabic also benefits from preprocessing steps such as reversing code-switching and emoji conversion [11].

Pre-trained multilingual models are integral to the NLP tasks for low-resource languages. BERT [3] itself offers two multilingual versions: cased and uncased. These models are impressive in their scope, being trained on over 100 languages. The training process leverages masked language modelling and next token prediction objectives, allowing the model to learn generalizable representations across languages. XLM [12] is another set of multilingual models that uses translation objective alongside causal and masked language modeling for pre-training. Similarly, mBART [13] builds upon the BART model [14] by using a multilingual pre-training objective. The objective is reconstructing the original text from a corrupted version in multiple languages, allowing mBART to develop robust denoising capabilities.

The growing popularity of cross-lingual transfer learning offers a promising approach to improve performance on Arabic NLP tasks. This is demonstrated by employing task-specific fine-tuning on English and French data to improve Arabic NLU performance [15]. Similarly, for abstractive summarization of Arabic text, fine-tuning multilingual models (mBERT and mBART) on Hungarian or English before fine-tuning again on Arabic data demonstrated performance gains [16]. These findings highlight the effectiveness of cross-lingual transfer learning in improving the performance of Arabic language processing tasks.

3. Methodology

In this section, we describe the methodology employed for detecting persuasion techniques in Arabic articles using a multilingual BERT model fine-tuned on English data.

3.1. Data Preparation

The data for this task was provided in the form of article files, with the corresponding labels given in a separate file. The label file contained information about the persuasion techniques used and the offsets indicating the span of text within the articles where these techniques were applied. There are 23 labels representing different persuasion techniques. These techniques are identified within the text at the token level, allowing for multi-label classification where each token can be associated with one or more techniques. This detailed annotation allows the model to recognize and classify multiple techniques within a single span of text.

For preprocessing, we first split the articles into paragraphs. This was done based on empty lines, effectively treating each paragraph as a separate instance. Once the articles were divided into paragraphs, we calculated the offsets for the persuasive spans within each paragraph. This allowed us to align the provided labels with the appropriate paragraphs.

3.2. Task Formulation

We formulated the task as a multi-class, multi-label token classification problem. This means that each token (or word) in the input text could be classified into one or more persuasion technique categories. This approach enables the model to recognize multiple techniques that may be present in a single span of text. After predicting labels for each of the tokens, consecutive tokens with the same labels define a span. Table 1 demonstrates an example.

Table 1

An input sequence of length 256, where each token can be assigned one or more persuasion techniques. Consecutive tokens with the same labels form a span. In this example, $\{x_3, x_4, x_5\}$ form a span for t_1; $\{x_1, x_2\}$ form a span for t_2; $\{x_5, x_6\}$ form a span for t_2; $\{x_4\}$ is a span for t_3; and $\{x_6\}$ is a span for t_3.

	x_1	x_2	x_3	x_4	x_5	x_6	...	x_{256}
t_1	0	0	1	1	1	0	0	0
t_2	1	1	0	0	1	1	0	0
t_3	0	0	0	1	0	1	0	0
:	:	:	:	:	:	:	:	:
t_{23}	0	0	0	0	0	0	0	0

3.3. Model and Training

We employed a multilingual BERT model for this task. Multilingual BERT (*mBERT*) is pre-trained on multiple languages, including Arabic and English, making it suitable for cross-lingual transfer learning. For the loss calculation, we used binary cross-entropy, which is well-suited for multi-label classification tasks.

Given the lack of Arabic training data and the zero-shot nature of the task for Arabic, we used the provided English training data to fine-tune the *mBERT* model. Since there was no Arabic data provided for validation, we utilized the Arabic validation dataset from the ArAIEval shared task on propaganda detection 2024.[1] This validation dataset consists of 921 documents, with an average of 30.25 tokens per document, and follows the same labelling and annotation guidelines. The following hyperparameters were used during training:

- Learning Rate: 5e-5
- Number of Epochs: 75
- Maximum Input Length: 256 tokens

Additionally, we utilized *pos_weights* to adjust the loss calculation. This helps in handling the class imbalance, ensuring that the model does not become biased towards the more frequent classes.

4. Results and Discussion

For evaluation, we used the modified F1-micro score, which accounts for partial matching of the spans. All the scores reported in this paper use that modified F1. Our model, fine-tuned using the English

[1] https://araieval.gitlab.io/task1/

training data and validated on the Arabic dev dataset, achieved an F1-micro score of 0.0998 on the dev set. When evaluated on the test set, the model's performance improved significantly, achieving an F1-micro score of 0.3009. The difference in performance between the dev and test sets could be attributed to the domain-specific nuances and potential distributional differences in the test set.

Below is a detailed breakdown of the F1-micro scores per technique on the validation set, as shown in Table 2.

Table 2
F1-micro scores per technique on the validation set.

Technique	F1-micro	Technique	F1-micro
Appeal to Values	0.6207	Questioning the Reputation	0.0000
Loaded Language	0.1855	Straw Man	0.2138
Consequential Oversimplification	0.6897	Repetition	0.0292
Causal Oversimplification	0.0542	Guilt by Association	0.0443
Appeal to Hypocrisy	0.0114	Conversation Killer	0.1724
False Dilemma-No Choice	0.0172	Whataboutism	0.2759
Slogans	0.0661	Obfuscation-Vagueness-Confusion	0.1034
Name Calling-Labeling	0.1257	Flag Waving	0.0709
Doubt	0.0483	Appeal to Fear-Prejudice	0.0472
Exaggeration-Minimisation	0.0983	Red Herring	0.5862
Appeal to Popularity	0.5517	Appeal to Authority	0.0949
Appeal to Time	0.8276		

The results reveal a significant variation in the model's performance across different persuasion techniques. Techniques such as Appeal to Time, Consequential Oversimplification, and Appeal to Values were detected more reliably, indicating that the model can effectively identify these patterns.

In contrast, techniques like Loaded Language, Straw Man, and Whataboutism showed moderate performance. Techniques like Questioning the Reputation, Repetition, False Dilemma-No Choice, and Appeal to Hypocrisy posed significant difficulties for the model. These techniques may be underrepresented in the training data, further complicating their detection.

The variation in performance can also be attributed to the nature and categorization of the techniques. Techniques that belong to the same category, such as different types of logical fallacies or emotional appeals, may share linguistic features that the model struggles to distinguish. For example, both Straw Man and Whataboutism involve misrepresentation or diversion tactics, which could confuse the model. On the other hand, techniques like Appeal to Values and Appeal to Popularity, which are more explicit and direct, tend to be easier for the model to identify.

It's important to note that no Arabic data was available for training. We relied on the English training data to fine-tune the multilingual BERT model. This cross-lingual transfer learning approach introduces additional challenges due to differences in linguistic structures and contextual usage between English and Arabic.

5. Conclusion and Future Work

With the increasing sophistication of persuasion techniques, particularly in Arabic-language content, it is crucial to focus research efforts on this area. This study investigated the effectiveness of a multilingual BERT model fine-tuned on English data for the task of Arabic persuasion detection. English was selected as the training language due to its extensive resources in Natural Language Processing (NLP) tasks, including propaganda detection. Our aim was to evaluate how these abundant resources could be leveraged to benefit languages with fewer resources, such as Arabic. This work achieved first place for Arabic on the leaderboard for the test set, demonstrating the potential of cross-lingual transfer learning [17]. However, there is still room for improvement.

Future work can explore how other high-resource languages impact the performance on Arabic. There might be various strategies to enhance the model's performance. Increasing the diversity and quantity of training data, particularly for techniques where performance was low, through data augmentation or the collection of additional labelled data, can help balance the dataset. Advanced fine-tuning techniques like focal loss can adjust the loss function to focus more on hard-to-classify examples, while dynamic sampling strategies can address class imbalance.

Additionally, incorporating more sophisticated features such as syntactic and semantic information, part-of-speech tags, or dependency parsing can provide the model with greater context and improve classification accuracy. Exploring alternative hidden layer representations within BERT may also yield better classification performance. By addressing these areas, future research can further improve the accuracy and robustness of models in detecting a wide range of persuasion techniques, ultimately enhancing their utility in real-world applications.

Acknowledgments

We acknowledge the assistance of the LT-Bridge Project (GA 952194) and DFKI for the use of their Virtual Laboratory. Also, authors have been supported financially by the EMLCT [2] programme during this entire work.

References

[1] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[2] J. Piskorski, N. Stefanovitch, F. Alam, R. Campos, D. Dimitrov, A. Jorge, S. Pollak, N. Ribin, Z. Fijavž, M. Hasanain, N. Guimarães, A. F. Pacheco, E. Sartori, P. Silvano, A. V. Zwitter, I. Koychev, N. Yu, P. Nakov, G. Da San Martino, Overview of the CLEF-2024 CheckThat! lab task 3 on persuasion techniques, in: [1], 2024.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[4] S. Yu, G. D. S. Martino, P. Nakov, Experiments in Detecting Persuasion Techniques in the News, 2019. URL: http://arxiv.org/abs/1911.06815, arXiv:1911.06815 [cs].

[5] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 70–98. URL: https://aclanthology.org/2021.semeval-1.7. doi:10.18653/v1/2021.semeval-1.7.

[6] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2343–2361. URL: https://aclanthology.org/2023.semeval-1.317. doi:10.18653/v1/2023.semeval-1.317.

[7] M. Hasanain, F. Alam, H. Mubarak, S. Abdaljalil, W. Zaghouani, P. Nakov, G. D. S. Martino, A. A. Freihat, Araieval shared task: Persuasion techniques and disinformation detection in arabic text, arXiv preprint arXiv:2311.03179 (2023).

[8] K. Gupta, D. Gautam, R. Mamidi, Volta at semeval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble (2021). URL: http://arxiv.org/abs/2106.00240, arXiv:2106.00240 [cs].

[9] K. Kaczyński, P. Przybyła, Homados at semeval-2021 task 6: Multi-task learning for propaganda detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, p. 1027–1031. URL: https://aclanthology.org/2021.semeval-1.141. doi:10.18653/v1/2021.semeval-1.141.

[2] https://mundus-web.coli.uni-saarland.de/

[10] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, H. Mubarak (Eds.), Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, European Language Resource Association, Marseille, France, 2020, pp. 9–15. URL: https://aclanthology.org/2020.osact-1.2.

[11] B. Tuck, F. Qachfar, D. Boumber, R. Verma, Detectiveredasers at araieval shared task: Leveraging transformer ensembles for arabic deception detection, in: Proceedings of ArabicNLP 2023, Association for Computational Linguistics, Singapore (Hybrid), 2023, p. 494–501. URL: https://aclanthology.org/2023.arabicnlp-1.45. doi:10.18653/v1/2023.arabicnlp-1.45.

[12] G. Lample, A. Conneau, Cross-lingual language model pretraining, arXiv preprint arXiv:1901.07291 (2019).

[13] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[15] K. Abboud, O. Golovneva, C. DiPersio, Cross-lingual transfer for low-resource arabic language understanding, in: H. Bouamor, H. Al-Khalifa, K. Darwish, O. Rambow, F. Bougares, A. Abdelali, N. Tomeh, S. Khalifa, W. Zaghouani (Eds.), Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, p. 225–237. URL: https://aclanthology.org/2022.wanlp-1.21. doi:10.18653/v1/2022.wanlp-1.21.

[16] M. Kahla, Z. G. Yang, A. Novák, Cross-lingual fine-tuning for abstractive arabic text summarization, in: Proceedings of the international conference on recent advances in natural language processing (ranlp 2021), 2021, pp. 655–663.

[17] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.