Itri Amigos at ArAIEval Shared Task: Transformer vs. Compression-Based Models for Persuasion Techniques and Disinformation Detection

Nouman Ahmed, Natalia Flechas Manrique, and **Jehad Oumer**
University of the Basque Country (UPV/EHU)
{anouman001, nflechas001, joumer001}@ikasle.ehu.eus

Abstract

Social media has significantly amplified the dissemination of misinformation. Researchers have employed natural language processing and machine learning techniques to identify and categorize false information on these platforms. While there is a well-established body of research on detecting fake news in English and Latin languages, the study of Arabic fake news detection remains limited. This paper describes the methods used to tackle the challenges of the ArAIEval shared Task 2023. We conducted experiments with both monolingual Arabic and multi-lingual pre-trained Language Models (LM). We found that the monolingual Arabic models outperformed in all four subtasks. Additionally, we explored a novel lossless compression method, which, while not surpassing pretrained LM performance, presents an intriguing avenue for future experimentation to achieve comparable results in a more efficient and rapid manner.

1 Introduction

The growing presence of social media as a way to quickly disseminate information to broad audiences, has had an undeniable shaping the sphere of public opinion. By their very nature, social media platforms have the associated peril of carrying messages that are erroneous at best, or carefully crafted to misinform and manipulate, at worst (e.g. Ishmuradova, 2019, Iida et al., 2022).

The development of NLP tools to fact check and explore persuasion techniques is a potential approach to counteract the effect of misinformation in social media. While this is an active area of research that is well established for English and other Latin languages, for Arabic news, there is still much room to explore. This paper describes the methodology used to tackle the classification tasks presented by the ArAIREval shared 2023 task (Hasanain et al., 2023), which builds upon WANLP 2022 (Alam et al., 2022). The tasks are described in Section 3.

We have mainly focused our efforts on two distinct approaches: on the one hand, the use of pretrained Language Models (LMs), which has been an established way to achieve state-of-the-art results in a range of NLP tasks (e.g. Devlin et al., 2018, Radford et al., 2019). Pre-trained LMs are advanced models, often based on Transformer architectures, that are pre-trained on massive datasets. On the other hand, we explore the approach presented by Jiang et al., 2023, which advocates for the use of simpler models that are less resource intensive and more interpretable. This method uses lossless compression and a distance metric with a k-nearest-neighbor classifier for text classification. The inconsistencies detected in this implementation will be detailed later on. The paper is organized as follows: Section 2 briefly talks about related work, Section 3 summarizes the datasets on each sub-task and Section 4 the methodology. Finally, Sections 5 and 6 present the conclusions and limitations, respectively.

All of the source code to reproduce the results is available in a Github repository [1].

2 Related Work

Prior research in the field of automated Arabic fake news detection predominantly relied on traditional machine learning classifiers, focusing mainly on binary classification scenarios. Mahlous and Allaith, 2021 applied NB, LR, SVM, RF, and XGB methods to classify Arabic news tweets as either fake or not. Among these, the Logistic Regression (LR) classifier achieved 87.8% accuracy using TF-IDF features at the n-grams level. Recent research has focused on assessing the performance of Transformer-based models. For example, Antoun et al., 2020 showed that AraBERT v02 achieved

[1] https://github.com/nouman-10/ArAIEval-Shared-Task/

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 543–548

high accuracy across various experimental scenarios. Similarly, Nassif et al., 2022 achieved favorable results on a Covid-19 fake news dataset using pre-trained models like RoBERTa-Base (Liu et al., 2019), ARBERT (Abdul-Mageed et al., 2021) and Arabic-BERT (Safaya et al., 2020). Alyoubi et al., 2023 show the good performance of MARABERT with CNNs for tweet classification.

While binary classification of news content has been a traditional approach, there is an emerging interest in multi-label classification scenarios. Several studies have ventured into this realm. Argotario, as introduced by Habernal et al., 2017, is a game-based platform designed to accumulate a dataset portraying a spectrum of fallacious arguments, with labels like: *ad hominem, appeal to emotion, red herring, hasty generalization, irrelevant authority*. In parallel Da San Martino et al., 2019 extracted and analyzed 451 articles sourced from 48 news outlets. These articles were annotated to highlight 18 unique propaganda techniques. These efforts emphasize the pivotal role of multi-label classification in revealing the nuanced tactics inherent in news narratives.

3 Sub-Tasks

Task 1, `Persuasion Technique Detection`, involves identifying persuasive elements within text snippets. Subtask A focuses on determining if a given multigenre snippet (composed of tweets and news paragraphs) contains content utilizing persuasion techniques, making it a binary classification task. Subtask B expands this by requiring further identification of specific propaganda techniques employed in the same multigenre snippet, turning the task into a multilabel classification.

Task 2, `Disinformation Detection`, centers around identifying and categorizing disinformation within tweets. Subtask 2A involves a binary classification task where the goal is to determine whether a tweet contains disinformation. Subtask 2B further refines this by requiring the detection of fine-grained disinformation classes, including *hate-speech, offensive content, rumors*, and *spam*.

4 Methodology

In this section, we describe our approach to processing the data, the models, and the experiments we conducted for all tasks.

4.1 Data Preparation and Preprocessing

During data preparation, we identified that in Subtasks 2A and 2B that a significant number of data points in these subtasks had the "text" feature set to the "NaN" (Not a Number) data type. Table 1 provides a breakdown of data points of all the sub-tasks including that lack of data in the "text" feature across the train and dev sets of Subtasks 2A and 2B. To address these anomalies, we converted "NaN" entries to strings. While we contemplated removing these anomalies from the dataset, the scoring system for the SharedTask mandated that all data points in the Dev set remain present and in their original sequence. Moreover, a clear class imbalance was identified at this stage, which we tried to tackle later by adding class weights to the model training.

Upon loading the data, we structured our experiments around three preprocessing settings: 1) Raw Data Processing: in this approach, no alterations were made. The text "feature" was used directly in its original form. 2) AraBERT Preprocessing: this method made use of the AraBERT preprocessing function. Key steps involved removing Arabic diacritic marks, stripping elongation characters and adding white spaces. Additionally, Hindi numerals were converted into their Arabic equivalents. 3) Link and Hashtag Removal: Building on the AraBERT prepossessing setting, this configuration further involved cleansing the text of "LINK" and "#" references. In Subsection 4.2, we detail the specific preprocessing configurations employed for our models across the various sub-tasks.

4.2 Our Approach

Our primary objective was to evaluate the efficacy of BERT-based models for persuasion and disinformation detection tasks. In this endeavor, we mainly examined AraBERT (Antoun et al.) [2]. AraBERT is an Arabic pretrained language model based on Google's BERT architecture (Devlin et al., 2018) with the BERT-Base configuration. The training dataset for AraBERT was curated from a myriad of sources, including OSCAR (Abadji et al., 2022), Arabic Wikipedia dump [3], and the 1.5B words Arabic Corpus (El-Khair, 2016) among others.

Beyond AraBERT, we experimented with models such as mBERT and XLM-RoBERTa (Conneau

[2] https://github.com/aub-mind/arabert/tree/master#AraBERT

[3] https://archive.org/details/arwiki-20190201

Sub-Task	Split	# Data Points	# NaN Data Points	Per Class Data Points
1A	Train	2427	0	'true': 1918, 'false': 509
1A	Dev	259	0	'true': 202, 'false': 57
1A	Test	503	0	'true': 331, 'false': 172
2A	Train	14147	21	'no-disinfo': 11491, 'disinfo': 2656
2A	Dev	2115	4	'no-disinfo': 1718, 'disinfo': 397
2A	Test	3729	0	'no-disinfo': 2853, 'disinfo': 876
2B	Train	2656	8	'HS': 1512, 'OFF': 500, 'SPAM': 453, 'Rumor': 191
2B	Dev	397	1	'HS': 226, 'OFF': 75, 'SPAM': 68, 'Rumor': 28
2B	Test	876	0	'HS': 442, 'SPAM': 241, 'OFF': 160, 'Rumor': 33

Table 1: Statistics of the data regarding all subtasks. Note that the number of data-points for sub-task 1B were the same as 1A but the dataset has too many classes to include here.

et al., 2019), with the aim of discerning the impact of multilingual data on our tasks. However, during the development phase, AraBERT consistently surpassed the performance of these multilingual models, likely due to its training on a substantial Arabic corpus. This observation aligns with studies like that of Alammary, 2022, emphasizing the efficacy of monolingual models in specific contexts. As a result, we opted for AraBERT.

Our secondary objective was to assess the performance of the model introduced by Jiang et al., 2023, who leveraged lossless compressors and the k-nearest-neighbor (kNN) algorithm for classification tasks. Their method is founded on the principle that lossless compressors (e.g., gzip, z2, lzma, and zstandard) are adept at representing regularities in data and that textual data within the same category share more similarities and regularities than those from distinct categories. By measuring the Normalized Compression Distance (NCD) between texts, this method capitalizes on the compression lengths to approximate the Kolmogorov complexity of data. This subsequently serves as the foundation for a distance metric used in kNN classification.

Substantial controversy has surfaced within the online research community concerning the work of Jiang et al., 2023. Prominent among these critiques are those from Sebastian Raschka[4] and Ken Schutte[5]. Both researchers highlighted potential discrepancies in the original paper's code and implementation. Specifically, they pinpointed an error in the kNN accuracy computation resulting from a flawed tie-breaking strategy, which may have inflated the reported results. Despite the critiques, Jiang et al.'s methodology offers a compelling approach to text classification. Seizing the opportunity presented by this shared task, we undertake an evaluation of the method through an independent implementation of the compressor-based classifier, adopting a different tie-breaking strategy for the kNN classifier. Our aim is to assess its performance on Arabic persuasion technique detection and disinformation detection tasks, and subsequently, to share these insights transparently with the research community.

4.3 Evaluation

In this section, we describe the results we achieved including the experimental setup.

4.3.1 Experimental Setup

In our experiments, we employed an updated version of AraBERT, which was trained on a substantially larger dataset, thus incorporating an expanded lexicon. The authors of the original model pinpointed a flaw in AraBERTv1's wordpiece vocabulary. The issue came from punctuation and numbers that were still attached to words when they trained the wordpiece vocab. They have since rectified this by introducing spaces around numerical digits and punctuation marks. To make sure this is compatible with any new downstream task, they have released a preprocessing function as well, that we apply in all our tasks before fine-tuning.

For the models, we chose to experiment with the three different versions of the base model. The first two models are trained on the same dataset but one (v2) uses pre-segmentation and the other (v02) does not. The last model (v02-Twitter) is trained on the combination of the same dataset plus 60M multi-dialect tweets from twitter as well. For all tasks, only the text data was used as a feature for training. To address the issue of class imbalance, class weights were computed and used during the training process. In addition to this, we experimented with removing hashtags and links, to see

[4] https://magazine.sebastianraschka.com/p/large-language-models-and-nearest

[5] https://kenschutte.com/gzip-knn-paper/

	Preprocessing		Task 1A		Task 1B		Task 2A		Task 2B	
Model	**+**	**++**	**Dev**	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Dev**	**Test**
AraBERT-v0.2	Yes	No	0.849	**0.755**	0.548	0.471	0.900	**0.904**	0.836	**0.828**
AraBERT-v2	Yes	No	0.861	0.748	0.598	0.550*	0.901	0.902	0.823	0.816
AraBERT-Twitter	Yes	No	0.868	0.747	0.538	0.481	0.909	0.898	**0.843**	0.817
AraBERT-v0.2	Yes	Yes	0.780	0.658	0.605	0.577	0.820	0.786	0.631	0.663
AraBERT-v2	Yes	Yes	**0.868**	0.749*	**0.606**	0.537	0.812	0.765	0.646	0.683
AraBERT-Twitter	Yes	Yes	0.779	0.658	0.601	**0.570**	**0.912**	0.898*	0.841	0.814*
gzip+knn (lowest-label-index)	No	No	0.803	0.658	0.499	0.393	0.800	0.772	0.687	0.713
gzip+knn (closer-neighbor)	No	No	0.745	0.636	0.489	0.345	0.830	0.801	0.664	0.681
gzip+knn (random-selection)	No	No	0.752	0.616	0.455	0.326	0.818	0.798	0.636	0.688
gzip+knn (k=3)	No	No	0.764	0.654	0.471	0.334	0.848	0.825	0.634	0.687
Majority baseline				0.658		0.360		0.765		0.505

Table 2: Results of AraBERT experiments on all Sub-Tasks. + and ++ denotes preprocessing using AraBERT preprocessor and removal of hashtags and LINKs respectively. Note that the results in bold are the models that performed the best but the models with $*$ are the ones that were submitted which may not align with the best score as some of the experiments were carried out after the deadline.

if they have a positive effect on the performance as well. All the models are trained for 10 epochs with the model performing best on validation set chosen for test evaluation. The learning rate and batch size was set to $2e-5$ and 16 respectively, with the model evaluated on the dev set after every epoch.

Alongside our submissions for the shared task using AraBERT pretrained models, we applied Jiang et al., 2023 approach to this specific shared task context. We utilized the gzip compressor for encoding the text data and calculated inter-textual distances using the Normalized Compression Distance (NCD). The k-Nearest Neighbors (kNN) classifier was employed with $k = 2$, mirroring the setup in Jiang et al. 2023's study.

For the kNN's tie-breaking mechanism, we evaluated three strategies: 1) Lowest-label-index: this method, which follows the convention employed in the original study, selects the label with the lowest index during a tie. 2) Random-selection: in instances of ties, this strategy randomly selects among the tied labels. 3) Closer-neighbor: this method gives preference to the label of the nearest tied neighbor. Furthermore, we conducted experiments using $k = 3$ for the kNN classifier, where tie-breaking mechanisms are inherently unnecessary due to the odd number of neighbors. In all subtasks, we opted for no preprocessing of the data, as our preliminary experiments revealed that preprocessing adversely affected the performance of the compression-based approach.

4.3.2 Results

Tasks 1A and 1B: Persuasion Technique Detection: Our submission with the AraBERT-v2 model recorded a Test score of 0.749 and 0.550, achieving 5th and 4th position in the leaderboard for the Task 1A and 1B respectively. Parallel to our primary experiments, our exploration into the methodology of Jiang et al., 2023 bore intriguing results. The compressor-based approach with the "lowest-label-index" tie-breaking strategy for the kNN classifier achieved a Test score of 0.658 in Task 1A, closely mirroring the majority baseline of 0.658. For Task 1B, the strategy performed above the baseline, achieving a score of 0.393 compared to the baseline of 0.360. It's noteworthy to mention that while the AraBERT models capitalized on their training over an expansive Arabic corpus, the gzip+knn approach showcased potential, particularly when considering its resource-efficient nature.

Tasks 2A and 2B: Disinformation Detection: The AraBERT-v02-Twitter displayed good performance, with Test scores of 0.898 and 0.814 for Tasks 2A and 2B respectively, achieving 8th and 7th position on the leaderboard. Meanwhile, the compressor-based classifier showed its merits once again. Using the "closer-neighbor" tie-breaking strategy, the gzip+knn approach produced a Test score of 0.801 for Task 2A, not far from the baseline of 0.765. In Task 2B, the "lowest-label-index"

strategy yielded a score of 0.713, surpassing the baseline of 0.505.

5 Conclusions

In our participation in the ArAIEval Shared Task, we predominantly employed Transformer-based models for persuasion techniques and disinformation detection tasks, given their demonstrated proficiency with Arabic textual data. Although our results highlighted the strengths of these models, we simultaneously recognized the emerging potential of the compression-based approach to text classification. While these compression-based methods are in their infancy, they offer exciting opportunities for continued research. Future studies should delve deeper into the applicability of lossless compressors for text classification and seek to identify non-parametric machine learning algorithms that best align with these compressors. Importantly, compressor-driven systems might be more advantageous in situations where resource efficiency and rapid processing take precedence over accuracy.

6 Limitations

While these experiments give us a promising avenue to explore in terms of detecting persuasion techniques and disinformation in Arabic text, even in a low-resource setting using compressors, there are a lot of limitations to these approaches. One thing to note is that although Pretrained LMs seem to recognize disinformation in these texts, there is no reliability to this score, as in order to fact-check any news, you need consolidating evidence to see if it is fake or not, rather than only looking at how it is worded. It can be argued that those instances in which the wording of a fake piece of news is indistinguishable from a truthful one are even more dangerous. To tackle this, ways to include other sources of data would help improve results.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ali Saleh Alammary. 2022. Bert models for arabic text classification: a systematic review. *Applied Sciences*, 12(11):5720.

Shatha Alyoubi, Manal Kalkatawi, and Felwa A. Abukhodair. 2023. The detection of fake news in arabic tweets using deep learning. *Applied Sciences*.

Wissam Antoun, Fady Baly, Rim Achour, A. Hussein, and Hazem M. Hajj. 2020. State of the art models for fake news detection tasks. *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 519–524.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics.

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*.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *ArXiv*, abs/1611.04033.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. *arXiv preprint arXiv:1707.06002*.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat. 2023. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Takeshi Iida, Jaehyun Song, José Luis Estrada, and Yuriko Takahashi. 2022. Fake news and its electoral consequences: a survey experiment on mexico. *AI & SOCIETY*.

Madinabonu Ishmuradova. 2019. Strategies for using fake news as a tool to manipulate public opinion.

Zhiying Jiang, Matthew Yang, Mikhail Tsirlin, Raphael Tang, Yiqin Dai, and Jimmy Lin. 2023. "low-resource" text classification: A parameter-free classification method with compressors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6810–6828, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ahmed Redha Mahlous and Ali Al-laith. 2021. Fake news detection in arabic tweets during the covid-19 pandemic. *International Journal of Advanced Computer Science and Applications*, 12.

Ali Bou Nassif, Ashraf Elnagar, Omar A. Elgendy, and Yaman Afadar. 2022. Arabic fake news detection based on deep contextualized embedding models. *Neural Computing & Applications*, 34:16019 – 16032.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.