PD-AR at ArAIEval Shared Task: A BERT-Centric Approach to Tackle Arabic Disinformation

Pritam Deka
Queen's University Belfast, UK
pdeka01@qub.ac.uk

Ashwathy T. Revi
University of Southampton, UK
atr1n17@soton.ac.uk

Abstract

This work explores Arabic disinformation identification, a crucial task in natural language processing, using a state-of-the-art NLP model. We highlight the performance of our system model against baseline models, including multilingual and Arabic-specific ones, and showcase the effectiveness of domain-specific pre-trained models. This work advocates for the adoption of tailored pre-trained models in NLP, emphasizing their significance in understanding diverse languages. By merging advanced NLP techniques with domain-specific pre-training, it advances Arabic disinformation identification.

1 Introduction

Disinformation is the deliberate creation and spreading of false or misleading information that can cause public harm or generate profit for organizations that participate in such practices (Tandoc Jr et al., 2018; de Cock Buning, 2018). The consequences of disinformation can be significant, affecting political decisions (Allcott and Gentzkow, 2017), manipulating public opinion, or even inciting violence. Detecting disinformation can be challenging because it can appear similar to real information and spread very quickly. Additionally, creators are constantly evolving their methods, making it more difficult to detect their content.

While disinformation detection in English has received much attention, the nuances, dialectical variations, and morphological richness of Arabic present unique challenges that have not been comprehensively addressed. The ArAlEval[1] shared Task 2: Disinformation Detection (Hasanain et al., 2023) aims to encourage further exploration of disinformation detection in Arabic content. It includes two sub-tasks: (A) to categorize whether a given tweet is disinformative, modelled as a binary classification task, and (B) detecting the fine-grained disinformation class for a tweet, modelled as a multiclass classification task with labels indicating the subtype of disinformation contained - hate speech, offensive, rumour or spam.

BERT-based (Devlin et al., 2018) models have been shown to be successful in understanding the context behind language and benefit from being able to transfer learned knowledge to various tasks. Due to these advantages, such models are very good in text classification tasks. We have, therefore, utilised BERT-based models for the shared task which has been pre-trained over Arabic text. We hypothesized that such pre-trained models will better understand Arabic text than a BERT model that has been pre-trained over English text. However, there are certain challenges when we are dealing with text that is code-mixed. Tweets usually contain texts that contain code-mixed text which may prove to be difficult to work with.

2 Related Work

Techniques used for disinformation detection include manually or automatically analyzing the content of a piece of information to identify features that are associated with disinformation, analyzing the social media activity around a piece of information to identify patterns that suggest it is being spread as disinformation and verifying the claims made in a piece of information using external knowledge (Hu et al., 2022a). In the domain of fake news detection, significant work has already been done which are covered by many seminal survey works on fake news detection such as (Shu et al., 2017; Oshikawa et al., 2018; Bondielli and Marcelloni, 2019; Elhadad et al., 2019; Zhou and Zafarani, 2020; Zhang and Ghorbani, 2020; Mridha et al., 2021; Hu et al., 2022b). Given that the surveys encompass research endeavors concerning fake news, encompassing both misinformation and disinformation detection, we shall employ these terms interchangeably within this section.

[1] https://gitlab.com/araieval/wanlp2023_araieval

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 570–575

However, with the advent of transformer models (Vaswani et al., 2017), the prospect of training neural network models on languages beyond English has become increasingly prominent. The application of transformer models to Arabic text is highly promising. These models, trained on extensive text data, can excel in Arabic NLP tasks. They can grasp sentiment nuances, crucial for sentiment analysis, and enhance translation accuracy in challenging Arabic-English translation tasks. For information retrieval, understanding Arabic query and document semantics is vital, where transformers show exceptional performance. This advancement has significantly improved Arabic NLP across domains. Consequently, a plethora of research studies focusing on Arabic fake news detection has emerged, many of which have been reviewed in prominent surveys like those by (Fouad et al., 2022; Nassif et al., 2022; Harrag and Djahli, 2022; Al-Yahya et al., 2021). Other related works also include shared task results on propaganda (Alam et al., 2022) and detection and reasoning of tweets (Mubarak et al., 2023).

The Covid-19 pandemic also led to a range of research works focusing on Arabic misinformation regarding the pandemic. The work by (Haouari et al., 2020) introduces a dataset for misinformation detection, covering various topical categories influenced by COVID-19, and presents benchmarking results for tweet-level verification. (Al-Rawi et al., 2022) examines the scale of Arabic COVID-19 disinformation, identifying prominent topics related to violations of civil liberties, vaccine-related conspiracies, and calls for action. (Ashraf et al., 2022) presents a machine learning-based system for detecting misinformation in Arabic tweets related to COVID-19 vaccination, achieving promising performance. The work by (Obeidat et al., 2022) introduces a comprehensive dataset annotated with fine-grained misinformation classes and situational information, and presents baseline results using various classifiers.

In contrast to the aforementioned work, this system paper investigates the effectiveness of a fine-tuned BERT model in binary and multi-class classification of disinformation work, thereby capturing a broader aspect of disinformation regarding Arabic Twitter data.

3 Data

For both the sub-tasks, we used the training and development sets for the competition since the test set labels were part of the competition. However, after the competition the test set labels were also released which is why in this paper we are including the details of the whole dataset for both the sub-tasks. For subtask 2A, the details of the dataset are shown in Table 1.

Dataset Details	disinfo	no-disinfo
train	2656	11491
dev	397	1718
test	876	2853
Overall	3929	16062

Table 1: Dataset details for subtask 2A

The columns disinfo and no-disinfo are the labels for the subtask where disinfo means having disinformation and no-disinfo means having no disinformation.

For the subtask 2B, the details of the dataset is shown in Table 2.

Dataset Details	HS	SPAM	OFF	Rumor
train	1512	453	500	191
dev	226	68	75	28
test	442	241	160	33
Overall	2180	762	735	252

Table 2: Dataset details for subtask 2B

Before training, we performed some preprocessing of the text using the Python RegEx[2] library as well as removal of NaN entries. The preprocessing steps include the removal of punctuation including symbols that includes both English and Arabic punctuation. We also normalized certain Arabic symbols, removal of repeating characters and hashtags, URLs and mentions.

4 System

In this section,we will first describe the system architecture and then discuss the implementation of the system.

4.1 System architecture

Our system is built upon the foundation of the AraBERTv0.2-Twitter model[3], which is an Arabic language model pre-trained on Arabic twitter

[2] https://docs.python.org/3/library/re.html

[3] https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter

specific text corpus, as described in (Antoun et al., 2021). The base model is a BERT model which has been pre-trained with Arabic text. To adapt this model to our specific tasks, we employ the model after extensive data preprocessing. All our experiments are conducted using the Huggingface framework (Wolf et al., 2020).

4.2 System implementation

Our workflow begins with text tokenization using the model's tokenizer. This crucial step breaks down the input text into its constituent tokens, ensuring compatibility with the model's architecture. Subsequently, we generate text embeddings using the model, creating numerical representations of the input text. For the training phase, we carefully select hyperparameters to optimize model performance. These hyperparameters include a maximum input length of 64 tokens, a batch size of 16, and a training duration of 20 epochs. The choice of a suitable learning rate is pivotal in fine-tuning, and we set it at 2e-5. To optimize model weights, we employ the Adam optimizer with an epsilon value of 1e-08 and a weight decay of 0.01. Throughout the training process, we periodically save model checkpoints, specifically after every 200 steps. This strategy allows us to monitor the model's progress and select the best-performing version based on accuracy for our subsequent tasks. Importantly, these hyperparameter settings draw inspiration from established works in the field, notably the work by (Devlin et al., 2018) and (Antoun et al., 2021). To effectively handle both subtask tasks, we employ different loss functions. For subtask 2A, a binary cross-entropy loss is used, as it aligns with the binary nature of this classification problem. In contrast, subtask 2B involves multi-class classification, thus necessitating the use of categorical cross-entropy loss. These choices of loss functions are made to suit the specific requirements and nature of each subtask.

5 Results

Since the test file with gold standard annotations was available for both sub-tasks, we have evaluated our model on this file. The results for both the sub-tasks are shown below in Table 3 and Table 4. For comparison, we have used the baseline approaches provided for the task. We also compared various other transformer based models with our system model for a fair comaprison. The baseline transformer models used are as follows: BERT-base (Devlin et al., 2018), XLM-RoBERTa-base (Conneau et al., 2019), RoBERTa-base (Liu et al., 2019), multlingual-BERT-base (Devlin et al., 2018) and CamelBERT (Inoue et al., 2021). To provide context, it's important to note that BERT-base and RoBERTa-base are transformer models pre-trained on English text. In contrast, XLM-RoBERTa-base and mBERT have undergone training on multilingual text, making them suitable for a broader range of languages. Lastly, the CamelBERT model has been pre-trained on Arabic text, rendering it particularly well-suited for the specific tasks this paper addresses. The performance comparison, as illustrated in the table, unequivocally underscores the superior capabilities of our system. Across both sub-tasks, our system consistently outperformed the baseline approaches as well as the other transformer-based models. We also report the top performing team results for a fair comparison with our results.

These results substantiate the efficacy of our approach, highlighting its robustness and suitability for the given tasks. The superior performance of our system showcases the importance of specialized pre-trained models, in enhancing the accuracy and effectiveness of domain specific natural language processing tasks.

Model	**Macro F-1**	**Micro F-1**
BERT-base-uncased	0.7921	0.8278
RoBERTa-base	0.4939	0.7758
XLM-RoBERTa-base	0.7618	0.8404
BERT-base-multilingual-uncased	0.8013	0.8696
BERT-base-arabic-camelbert-mix	0.8428	0.8924
Task Baseline (Random)	0.4763	0.5154
Task Baseline (Majority)	0.4335	0.7651
Top Team (DetectiveRedasers)	**0.8626**	**0.9048**
Our system	0.8595	0.9021

Table 3: Macro and micro f-1 comparison for subtask 2A

Model	**Macro F-1**	**Micro F-1**
BERT-base-uncased	0.4856	0.7271
RoBERTa-base	0.3905	0.6872
XLM-RoBERTa-base	0.4287	0.7431
BERT-base-multilingual-uncased	0.6303	0.7659
BERT-base-arabic-camelbert-mix	0.6809	0.8002
Task Baseline (Random)	0.2243	0.2603
Task Baseline (Majority)	0.1677	0.5046
Top Team (DetectiveRedasers)	**0.7541**	**0.8356**
Our system	0.7209	0.8174

Table 4: Macro and micro f-1 comparison for subtask 2B

6 Discussion

In this section, we delve into the discussion of the results obtained from our experiments with various BERT-based models. Our results, as illustrated in Table 3 and Table 4, offer valuable insights into the effectiveness of different pre-trained BERT-based models. Notably, we observed that BERT-based models specifically pre-trained on Arabic text consistently outperformed their generic and multilingual counterparts. This observation underscores the importance of leveraging language-specific pre-trained models when working with Arabic language data. Furthermore, our experiments revealed an intriguing finding regarding the role of training data sources. Specifically, we noted that a BERT model pre-trained on Arabic Twitter data exhibited superior performance compared to models trained on more general Arabic text. This outcome suggests that the unique characteristics of Twitter data, such as the distinctive writing style shaped by the platform's character limitations, can be harnessed to enhance the performance of NLP models for tasks involving Twitter content. It is worth highlighting that while the CamelBERT model has been trained on Arabic text, the Twitter-specific Arabic BERT model that we opted for our work showed better performance. This preference demonstrates that, even within the domain of Arabic language, domain-specific pre-trained models can offer advantages over more generalized alternatives. In essence, our findings emphasize the significance of tailoring pre-training data to the specific characteristics and requirements of the target task.

6.1 Ablation Study

As part of our discussion, we also did an ablation study wherein we experimented with our model for the multi-class task by dropping some of the classes. Based on the class instances, we first drop the Rumor class since it has the least number of instances across the train, dev and test sets. We then proceeded with the same experiments whose details are presented in the Table 5.

Comparing Table 5 with Table 4, we can see that there is an increase in the macro as well as micro f-1 scores across all the models. One reason for this could be the class imbalance in the dataset. Across the whole dataset, the rumor class has the lowest number of instances. Therefore, removing those instances may lead to a more balanced dataset thereby increasing the model performance. However, in order to verify this, we experimented by keeping the Rumor class and dropping a different class, SPAM which has a higher number of instances than Rumor and has a similar number of instances with the OFF class. The results of this experiment is shown in Table 6.

Model	Macro f-1	Micro f-1
BERT-base-uncased	0.6459	0.7663
RoBERTa-base	0.4963	0.6856
XLM-roBERTa-base	0.5811	0.7746
BERT-base-multilingual-uncased	0.7150	0.7781
BERT-base-arabic-camelbert-mix	0.7331	0.8173
Our system	0.7926	0.8505

Table 5: Results without the Rumor class

Model	Macro f-1	Micro f-1
BERT-base-uncased	0.3237	0.6881
RoBERTa-base	0.2736	0.6960
XLM-roBERTa-base	0.2823	0.6992
BERT-base-multilingual-uncased	0.5424	0.7102
BERT-base-arabic-camelbert-mix	0.5604	0.7370
Our system	0.6373	0.7574

Table 6: Results without the SPAM class

We can see from Table 6 that dropping the SPAM class and keeping the Rumor class leads to a decrease in model performance across all models for both macro as well as micro f-1. This shows that there is an imbalance in the dataset with the low instances of the Rumor class. In order to mitigate this issue, one way would be to increase the number of instances while data collection and the other would be to make use of data augmentation synthetically and append the new synthetic data to the dataset. However, although the data augmentation seems like a viable option without having to collect new data, further research is required in order to find suitable augmentation methods that can improve the performance of the model without generating noise and bias.

7 Conclusion

In this study, we have presented a comprehensive analysis of our system's performance in addressing the task of Arabic disinformation. The results of our evaluation unequivocally illustrate the superiority of our system over various baseline approaches, including those based on generic and multilingual transformer models. Notably, our system's outstanding performance in both sub-tasks underscores the significance of such language-specific

pre-trained models in enhancing the precision and utility of natural language processing applications.

Furthermore, the superiority of our system, even when compared to CamelBERT, a model pre-trained on Arabic text, highlights the importance of considering the specific nuances of data sources. In our case, a pre-trained model on Twitter-specific Arabic text data proved to be an advantageous choice, particularly for tasks involving Twitter data, where the writing style is distinct due to character limitations.

References

Ahmed Al-Rawi, Abdelrahman Fakida, Kelly Grounds, et al. 2022. Investigation of covid-19 misinformation in arabic on twitter: Content analysis. *Jmir Infodemiology*, 2(2):e37007.

Maha Al-Yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and Amr Essam. 2021. Arabic fake news detection: comparative study of neural networks and transformer-based approaches. *Complexity*, 2021:1–10.

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.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding.

Nsrin Ashraf, Hamada Nayel, and Mohamed Taha. 2022. Misinformation detection in arabic tweets: A case study about covid-19 vaccination. *Benha Journal of Applied Sciences*, 7(5):265–268.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

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. *arXiv preprint arXiv:1911.02116*.

Madeleine de Cock Buning. 2018. *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union.

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*.

Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2019. Fake news detection on social media: a systematic survey. In *2019 IEEE Pacific Rim conference on communications, computers and signal processing (PACRIM)*, pages 1–8. IEEE.

Khaled M Fouad, Sahar F Sabbeh, and Walaa Medhat. 2022. Arabic fake news detection using deep learning. *Computers, Materials & Continua*, 71(2).

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*.

Fouzi Harrag and Mohamed Khalil Djahli. 2022. Arabic fake news detection: A fact checking based deep learning approach. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–34.

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.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022a. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

LinMei Hu, SiQi Wei, Ziwang Zhao, and Bin Wu. 2022b. Deep learning for fake news detection: A comprehensive survey. *AI Open*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). 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. *arXiv preprint arXiv:1907.11692*.

Muhammad F Mridha, Ashfia Jannat Keya, Md Abdul Hamid, Muhammad Mostafa Monowar, and Md Saifur Rahman. 2021. A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9:156151–156170.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *arXiv preprint arXiv:2305.04927*.

Ali Bou Nassif, Ashraf Elnagar, Omar Elgendy, and Yaman Afadar. 2022. Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, 34(18):16019–16032.

Rasha Obeidat, Maram Gharaibeh, Malak Abdullah, and Yara Alharahsheh. 2022. Multi-label multi-class covid-19 arabic twitter dataset with fine-grained misinformation and situational information annotations. *PeerJ Computer Science*, 8:e1151.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.