UL & UM6P at ArAIEval Shared Task: Transformer-based Model for Persuasion Techniques and Disinformation Detection in Arabic

Salima Lamsiyah[1], **Abdelkader El Mahdaouy**[2], **Hamza Alami**[2]
Ismail Berrada[2] and **Christoph Schommer**[1]
[1]Dept. of Computer Science, Faculty of Science, Technology and Medicine,
University of Luxembourg, Luxembourg
[2]College of Computing, Mohammed VI Polytechnic University, Morocco
{firstname.lastname}@{uni.lu[1], um6p.ma[2]}

Abstract

In this paper, we introduce our participating system to the ArAIEval Shared Task, addressing both the detection of persuasion techniques and disinformation tasks. Our proposed system employs a pre-trained transformer-based language model for Arabic, alongside a classifier. We have assessed the performance of three Arabic Pre-trained Language Models (PLMs) for sentence encoding. Additionally, to enhance our model's performance, we have explored various training objectives, including Cross-Entropy loss, regularized Mixup loss, asymmetric multi-label loss, and Focal Tversky loss. On the official test set, our system has achieved micro-F1 scores of 0.7515, 0.5666, 0.904, and 0.8333 for Sub-Task 1A, Sub-Task 1B, Sub-Task 2A, and Sub-Task 2B, respectively. Furthermore, our system has secured the 4th, 1st, 3rd, and 2nd positions, respectively, among all participating systems in sub-tasks 1A, 1B, 2A, and 2B of the ArAIEval shared task.

1 Introduction

Social media platforms have transformed into significant spaces where people communicate and collect information from various sources. However, along with this positive shift, a significant amount of false, misleading, and harmful content has also emerged. This includes various forms of problematic content like misinformation, disinformation, and malinformation in the form of spreading propaganda, conspiracy theories, rumors, hoaxes, fake news, false statements, hate speech, cyberbullying, and among others (Oshikawa et al., 2020; Alam et al., 2021; Sharara et al., 2022; Essefar et al., 2021; Nakov et al., 2021a; Alam et al., 2022; Lamsiyah et al., 2023; Mubarak et al., 2023).

Furthermore, the surge in online communication platforms has also made it more important to understand how people try to persuade each other. The persuasion detection task involves the identification and analysis of communication strategies aimed at influencing individuals' beliefs or actions. It encompasses recognizing techniques such as emotional appeals, logical reasoning, and rhetorical devices in various forms of content (Dimitrov et al., 2021). Propaganda, a subset of persuasive communication, refers to the deliberate dissemination of information, often with a biased or misleading intent, to manipulate opinions or behaviors. It involves employing well-defined psychological and rhetorical methods to sway audiences (Alam et al., 2022). Several shared tasks have been organized for the detection of propaganda techniques in text and memes. This includes the NLP4IF-2019 shared task on Fine-Grained Propaganda Detection (Da San Martino et al., 2019), SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles (Da San Martino et al., 2020), and SemEval-2021 task 6 on Detection of Persuasion Techniques in Texts and Images (Dimitrov et al., 2021). In addition to detecting propaganda techniques, another intriguing task is to identify misleading content within social media. This aims to uncover various forms of disinformation, such as hate speech, offensive language, rumors, and spam (Barrón-Cedeno et al., 2020; Nakov et al., 2021b; Shahi et al., 2021).

Most of the previously mentioned research works have primarily focused on the English language. Therefore, there is a noteworthy need to develop such methods for the Arabic language, which is spoken by a considerable number of people globally, with an estimated 372 to 446 million speakers worldwide. With the aim of bridging this language gap, Hasanain et al. (2020) have presented a description of three Arabic tasks that were offered as part of the third edition of the CheckThat! lab at CLEF 2020. It focused on false information propagated on Arabic social media, particularly on Twitter. Furthermore, Alam et al. (2022) have run a shared task on detecting propaganda techniques in Arabic tweets as part of the WANLP 2022 work-

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 558–564

shop. More recently, Hasanain et al. (2023) have introduced the ArAIEval shared task that includes two tasks: (i) persuasion techniques detection (Sub-Task 1A and Sub-Task 1B), and (ii) disinformation detection (Sub-Task 2A and Sub-Task 2B) in the Arabic Language.

In this paper, we present our submitted system for the ArAIEval shared task (Hasanain et al., 2023), where we tackle both the tasks of detecting persuasion techniques and identifying disinformation. Our system utilizes a deep learning model that comprises a transformer-based Pre-trained Language Model (PLM) encoder designed for the Arabic language, coupled with a classifier. The classifier consists of a dropout layer followed by a linear layer. To encode text inputs, we have evaluated the performance of three Arabic PLMs: ARBERTv2, MARBERTv2, and AraBERT-large (Abdul-Mageed et al., 2021; Elmadany et al., 2022; Antoun et al., 2020). During the model training process, we have explored the following training objectives:

- **Sub-Task 1A and Sub-Task 2A**: We have used the cross-entropy loss and the regularized Mixup (RegMixup) loss (Pinto et al., 2023).
- **Sub-Task 1B**: We have evaluated the binary cross-entropy loss, the asymmetric loss for multi-label classification (Ben-Baruch et al., 2020), and the RegMixup loss (Pinto et al., 2023).
- **Sub-Task 2B**: We have employed the cross-entropy loss and the Focal Tversky loss (Abraham and Khan, 2018).

Our system is evaluated using the weighted-average Precision and Recall as well as the micro and macro F1 score. It has achieved micro-F1 scores of 0.7515, 0.5666, 0.904, and 0.8333 on the test sets of Sub-Task 1A, Sub-Task 1B, Sub-Task 2A, and Sub-Task 2B, respectively. Furthermore, our system has secured the 4th, 1st, 3rd, and 2nd positions, respectively, among all participating systems in the corresponding Sub-Tasks of the ArAIEval shared task. It is worth mentioning that the best results have been obtained using the ARBERT sentence encoder for both Sub-Task 1B and Sub-Task 2A. While, for Sub-Task 1A and Sub-Task 2B, the best performance has been achieved using the MARBERTv2 encoder.

2 Data

The ArAIEval shared task (Hasanain et al., 2023) comprises two tasks: persuasion techniques detecting (Sub-Task 1A and Sub-Task 1B), as well as disinformation detection (Sub-Task 2A and Sub-Task 2B) in Arabic. Table 1 describes the provided data for each sub-task. For persuasion techniques detection, the ArAIEval organizers propose the following two sub-tasks:

- **Sub-Task 1A**: is a binary classification task that detects whether a given input tweet or news paragraph contains a persuasion technique.
- **Sub-Task 1B**: is a multi-label classification task that aims to identify the persuasion techniques in a given tweet or news paragraph. The label set of this sub-task contains 24 labels.

For disinformation detection, the ArAIEval organizers provide data for the following two sub-tasks:

- **Sub-Task 2A**: is a binary classification task that aims to detect whether a given input tweet is disinformative.
- **Sub-Task 2B**: is a multi-class classification task that aims to identify the disinformation class of a given input tweet. The class labels include hate-speech, offensive, rumor, and spam.

Task	Train Set	Dev Set	Test Set	Num of classes	Domain
Sub-Task 1A	2427	259	503	2	Twitter and News
Sub-Task 1B	2427	259	503	24	Twitter and News
Sub-Task 2A	14147	2115	3729	2	Twitter
Sub-Task 2B	2656	397	876	4	Twitter

Table 1: ArAIEval subtasks data description

3 System Overview

3.1 Model Architecture

The proposed system comprises a BERT-based Arabic PLM encoder and a single classifier. The classifier consists of a dropout layer followed by a linear layer (feed-forward layer) with an activation function. The number of output units in the linear layer matches the number of classes. For Sub-Task 1A, 2A, and 2B, we have employed the Softmax activation, while for Sub-Task 1B, we have used the

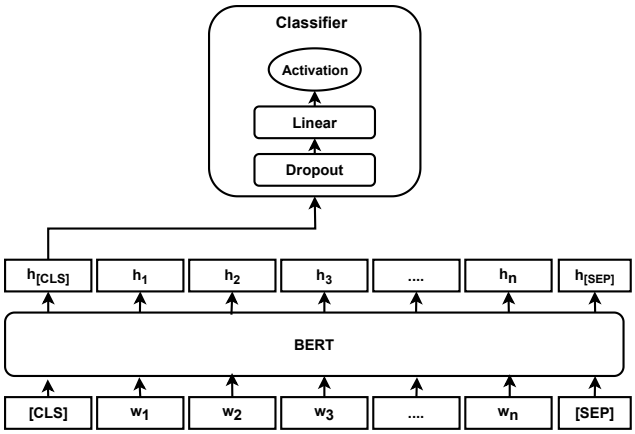

Figure 1: Overall Model Architecture

Sigmoid activation. The overall model architecture is depicted in Figure 1.

For the input texts encoding, we have explored the performance of three existing BERT-based Arabic PLMs, including ARBERTv2, MARBERTv2, and AraBERT-large(Abdul-Mageed et al., 2021; Elmadany et al., 2022; Antoun et al., 2020). These PLMs have been trained on large Arabic textual corpora, covering both Modern Standard Arabic and Dialectal Arabic, using the masked language modeling objective function.

As shown in Figure 1, given an input text of length m, the PLM's tokenizer split it into n sub-words and append the $[CLS]$ and $[SEP]$ special tokens, representing the start and end of the input sequence, to the tokenized text $([CLS], w_1, w_2, w_3, ..., w_n, [SEP])$. Then, the BERT-based encoder is fed with the tokenized text and outputs the contextualized word embedding $h_{[CLS]}, h_1, h_2, h_3, ..., h_n, h_{[SEP]}$. Finally, the pooled output of the $[CLS]$ token is passed to the classifier to predict the class label of the input text.

3.2 Training objectives

For model training, we have explored the following training objectives:

- $\mathcal{L}_{CE}$ denotes the Cross-Entropy (CE) loss;
- $\mathcal{L}_{BCE}$ denotes the Binary Cross-Entropy (BCE) loss;
- $\mathcal{L}_{ASL}$ denotes the Asymmetric Loss (ASL) for multi-label classification (Ben-Baruch et al., 2020). This loss function deals with the negative-positive imbalance in multi-label classification;
- $\mathcal{L}_{FT}$ denotes the Focal Tversky (FT) loss (Abraham and Khan, 2018). This loss function is a generalization of the focal loss and employs the Tversky index. It deals with the class imbalance problem.
- $\mathcal{L}_{RegMix}$ denotes the Regularized Mixup (RegMix) loss (Pinto et al., 2023). This loss is employed as a regularizer to the cross-entropy loss to improve the model's generalization. Formally, give two pair of examples and their corresponding labels from the training dataset (x_i, y_i) and (x_j, y_j), the Mixup is calculated as $\tilde{x_i} = \lambda \cdot x_i + (1 - \lambda) \cdot x_j$ and $\tilde{y_i} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j$. Where $\lambda \sim Beta(\alpha, \alpha) \in [0, 1]$ for $\alpha \in [0, \infty[$. Then, the RegMix loss is calculated as follows:

$$\mathcal{L}^{*}_{RegMix} = \mathcal{L}_{*}(x, y) + p \cdot \mathcal{L}_{*}(\tilde{x}, \tilde{y}) \quad (1)$$

  where $*$ and p denote a loss function like cross-entropy loss and Mixup weighting hyper-parameter. Since text mixup is not feasible, we employ mixup of the pooled output (h_i and h_j) of x_i and x_j.

For our models training on each sub-task, we have investigated the following training objectives:

- **Sub-Task 1A and Sub-Task 2A**: $\mathcal{L}_{CE}$ and $\mathcal{L}^{CE}_{RegMix}$
- **Sub-Task 1B**: $\mathcal{L}_{BCE}$, $\mathcal{L}_{ASL}$, and $\mathcal{L}^{ASL}_{RegMix}$
- **Sub-Task 2B**: $\mathcal{L}_{CE}$ and $\mathcal{L}_{FT}$

4 Experiments and Results

In this section, we present the experiment settings and the obtained results for each sub-task.

4.1 Experiment Settings

All our models have been implemented using the Pytorch deep learning framework, Pytorch Lightning, and Hugging Face Transformers library. We have performed our experiments on a Dell PowerEdge C4140 server, having 4 Nvidia V100 SXM2 32GB. For all sub-tasks, we have trained our models for a maximum of 10 epochs with a batch size of 16 examples and a learning rate of 1×10^{-5}. Early stopping is configured to 3 epochs. Besides, a weight decay of 1×10^{-3} is applied to all the layers of the model weights except biases and Layer Normalization (LayerNorm). In all our experiments,

		Dev				Test			
Encoder	**Loss**	**Precision**	**Recall**	**F1-macro**	**F1-micro**	**Precision**	**Recall**	**F1-macro**	**F1-micro**
AraBERT	$\mathcal{L}_{CE}$	0.826	0.834	0.7432	0.834	0.7438	0.7416	0.7152	0.7416
ARBERTv2		0.8102	0.8147	0.723	0.8147	0.7494	0.7455	0.721	0.7455
MARBERTv2		0.8437	0.8494	0.7703	0.8494	0.7569	0.7495	0.7281	0.7495
AraBERT	$\mathcal{L}^{CE}_{RegMix}$	0.8452	0.8533	0.7489	0.8533	0.7409	0.7475	0.7085	0.7475
ARBERTv2		0.8122	0.8263	0.6833	0.8263	0.7259	0.7356	0.6847	0.7356
MARBERTv2		**0.8622**	**0.8687**	**0.7893**	**0.8687**	**0.7476**	**0.7515**	**0.7186**	**0.7515**†

Table 2: The obtained results of our system on Sub-Task 1A. Our official submission results are highlighted in bold font. † is attached to the best obtained micro-F1 score.

		Dev				Test			
Encoder	**Loss**	**Precision**	**Recall**	**F1-macro**	**F1-micro**	**Precision**	**Recall**	**F1-macro**	**F1-micro**
AraBERT	$\mathcal{L}_{BCE}$	0.5757	0.5286	0.1166	0.6175	0.5181	0.4574	0.1035	0.5427
ARBERTv2		0.5879	0.5207	0.1176	0.619	0.527	0.4695	0.1044	0.5546
MARBERTv2		0.5397	0.5247	0.1098	0.6011	0.4808	0.4585	0.0976	0.5401
AraBERT	$\mathcal{L}_{ASL}$	**0.6286**	**0.6864**	**0.3296**	**0.6622**	**0.5833**	**0.5415**	**0.2156**	**0.5666**
ARBERTv2		0.592	0.6568	0.3315	0.6201	0.56	0.5526	0.2242	0.5538
MARBERTv2		0.6206	0.6844	0.2971	0.6438	0.5578	0.5604	0.1908	0.5766†
AraBERT	$\mathcal{L}^{ASL}_{RegMix}$	0.6059	0.6726	0.3285	0.644	0.5747	0.5482	0.2286	0.5651
ARBERTv2		0.5819	0.6785	0.3168	0.6243	0.5555	0.5637	0.2064	0.5678
MARBERTv2		0.6124	0.6903	0.2966	0.6512	0.5809	0.5648	0.2082	0.5756

Table 3: The obtained results of our system on Sub-Task 1B. Our official submission results are highlighted in bold font. † is attached to the best obtained micro-F1 score.

		Dev				Test			
Encoder	**Loss**	**Precision**	**Recall**	**F1-macro**	**F1-micro**	**Precision**	**Recall**	**F1-macro**	**F1-micro**
AraBERT	$\mathcal{L}_{CE}$	**0.9118**	**0.9144**	**0.8535**	**0.9144**	**0.9028**	**0.904**	**0.8645**	**0.904**
ARBERTv2		0.8972	0.9012	0.8283	0.9012	0.895	0.8976	0.8521	0.8976
MARBERTv2		0.9064	0.9078	0.8463	0.9078	0.905	0.9067	0.8672	0.9067†
AraBERT	$\mathcal{L}^{CE}_{RegMix}$	0.9101	0.9125	0.8515	0.9125	0.9034	0.9037	0.8656	0.9037
ARBERTv2		0.9002	0.9045	0.8294	0.9045	0.8935	0.8965	0.8479	0.8965
MARBERTv2		0.9096	0.913	0.845	0.913	0.9016	0.904	0.8583	0.904

Table 4: The obtained results of our system on Sub-Task 2A. Our official submission results are highlighted in bold font. † is attached to the best obtained micro-F1 score.

		Dev				Test			
Encoder	**Loss**	**Precision**	**Recall**	**F1-macro**	**F1-micro**	**Precision**	**Recall**	**F1-macro**	**F1-micro**
AraBERT	$\mathcal{L}_{CE}$	0.8234	0.8262	0.7973	0.8262	0.8205	0.8242	0.724	0.8242
ARBERTv2		0.8261	0.8287	0.8109	0.8287	0.8174	0.8151	0.7336	0.8151
MARBERTv2		0.8327	0.8363	0.795	0.8363	0.8345	0.8379	0.7443	0.8379†
AraBERT	$\mathcal{L}_{FT}$	0.8182	0.8212	0.7898	0.8212	0.818	0.8208	0.7231	0.8208
ARBERTv2		0.835	0.8388	0.8	0.8388	0.8055	0.8071	0.7024	0.8071
MARBERTv2		**0.8471**	**0.8514**	**0.8121**	**0.8514**	**0.8367**	**0.8333**	**0.7388**	**0.8333**

Table 5: The obtained results of our system on Sub-Task 2B. Our official submission results are highlighted in bold font. † is attached to the best obtained micro-F1 score.

we have fixed the maximum sequence length to 128. The hyper-parameters α (Beta distribution parameter) and p of the $\mathcal{L}_{RegMix}$ are set to 20 and 0.2, respectively. For $\mathcal{L}_{ASL}$ loss function, the hyper-parameters γ_- and γ_+ are fixed to 4 and 1, respectively. The hyper-parameter α of the Focal Tversky loss ($\mathcal{L}_{FT}$) is set to 0.5. It is worth mentioning that we have trained, validated, and evaluated our models on the officially provided splits for training, validation, and development, respectively. For the evaluation purpose, we have employed the weighted Recall and Precision as well as the micro and macro F1 scores.

4.2 Results

4.2.1 Sub-Task 1A

Table 2 summarizes our obtained results for Sub-Task 1A. The overall results show that employing the MARBERTv2 encoder leads to better performance using both the cross-entropy loss and the RegMix loss. Although the RegMix training objective largely enhances the results on the dev set, it achieves small performance improvements on the test set when AraBERT and MARBERTv2 encoders are utilized. The best results are obtained using the RegMix training objective in conjunction with the MARBERTv2 encoder. The latter corresponds to our official submission.

4.2.2 Sub-Task 1B

Table 3 shows our system's obtained results for Sub-Task 1B. The overall results demonstrate that the AraBERT and MARBERTv2 lead to better results for most training objectives. The asymmetric loss ($\mathcal{L}_{ASL}$) improves the classification results of all the used encoders and shows important performance increments for the macro-F1 and micro-F1 scores. Besides, the best micro-F1 score on the test set is obtained using the asymmetric loss in conjunction with the MARBERTv2 encoder. The RegMix training objective with the ASL loss ($\mathcal{L}^{ASL}_{RegMix}$) enhances the results when the ARBERTv2 encoder is employed. However, it negatively impacts the performance when the other two encoders are utilized. For the official evaluation, we have submitted our model that uses an AraBERT encoder, trained using the ASL loss.

4.2.3 Sub-Task 2A

Table 4 presents our obtained results for Sub-Task 2A. The overall results show that AraBERT and MARBERTv2 encoders yield better results than ARBERTv2. The RegMix training objective slightly degrades the F1 scores performance of our systems. Our best micro-F1 score is obtained using MARBERTv2 in conjunction with the CE training objective. Whereas, our official submitted model is trained using the CE loss and AraBERT encoder.

4.2.4 Sub-Task 2B

Table 5 summarizes our obtained results for Sub-Task 2B. The overall results show that the MARBERTv2 outperforms the other pre-trained models. Although the Focal Tversky loss has been shown to improve the results of ARBERTv2 and MARBERTv2 on the dev set, it negatively impacts our model performance on the test set. The best micro-F1 score is achieved by using the MARBERTv2 encoder in conjunction with CE loss. Whereas, our official submitted model is trained using the FT loss and MARBERTv2 encoder.

5 Discussion

The obtained results have shown that the training objective and the text encoder have a significant impact on our models' performance. The overall results demonstrate the effectiveness of the PLMs encoders that are pre-trained on large text corpora from the same domain as the target downstream tasks (MARBERTv2). A straightforward path of future research work is to investigate the performances of other state-of-the-art Arabic PLMs and other training objectives that deal with the class imbalance problem.

6 Conclusion

In this paper, we have introduced our submitted system to the ArAIEval Shared Task for persuasion techniques and disinformation detection in Arabic. Our System uses a deep learning model that consists of a transformer-based Pre-trained Language Model (PLM) encoder for the Arabic language and a classifier. For the model training, we have explored several training objectives and assessed the performance of three Arabic PLMs. On the official test set, our system has obtained micro-F1 scores of 0.7515, 0.5666, 0.904, and 0.8333 for Sub-Task 1A, Sub-Task 1B, Sub-Task 2A, and Sub-Task 2B, respectively. Besides, it has been ranked in the 4th, 1st, 3rd, and 2nd positions among all participating systems in Sub-Task 1A, Sub-Task 1B, Sub-Task 2A, and Sub-Task 2B, respectively.

References

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.

Nabila Abraham and Naimul Mefraz Khan. 2018. A novel focal tversky loss function with improved attention u-net for lesion segmentation.

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 The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. 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.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 215–236. Springer.

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2020. Asymmetric loss for multi-label classification.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2022. Orca: A challenging benchmark for arabic language understanding. *arXiv preprint arXiv:2212.10758*.

Kabil Essefar, Abdellah El Mekki, Abdelkader El Mahdaouy, Nabil El Mamoun, and Ismail Berrada. 2021. CS-UM6P at SemEval-2021 task 7: Deep multi-task learning model for detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1135–1140, Online. Association for Computational Linguistics.

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.

Maram Hasanain, Fatima Haouari, Reem Suwaileh, Zien Sheikh Ali, Bayan Hamdan, Tamer Elsayed, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of checkthat! 2020i arabic: Automatic identification and verification of claims in social media. In *Conference and Labs of the Evaluation Forum*.

Salima Lamsiyah, Abdelkader El Mahdaouy, Hamza Alami, Ismail Berrada, and Christoph Schommer. 2023. UL & UM6P at SemEval-2023 task 10: Semi-supervised multi-task learning for explainable detection of online sexism. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 644–650, Toronto, Canada. Association for Computational Linguistics.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and identifying the reasons for deleted tweets before they are posted. *Frontiers in Artificial Intelligence*, 6.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021a. Covid-19 in bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021b. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 639–649. Springer.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. 2023. Regmixup: Mixup as a regularizer can surprisingly improve accuracy and out distribution robustness.

Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. Overview of the clef-2021 checkthat! lab: Task 3 on fake news detection. In *CLEF (Working Notes)*, pages 406–423.

Mohamad Sharara, Wissam Mohamad, Ralph Tawil, Ralph Chobok, Wolf Assi, and Antonio Tannoury. 2022. Arabert model for propaganda detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 520–523.