ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text

Maram Hasanain[1], **Firoj Alam**[1], **Hamdy Mubarak**[1], **Samir Abdaljalil**[1],
Wajdi Zaghouani[2], **Preslav Nakov**[3],
Giovanni Da San Martino[4], **Abed Alhakim Freihat**[5]
[1]Qatar Computing Research Institute, HBKU, Qatar
[2]Hamad Bin Khalifa University, Qatar,
[3]Mohamed bin Zayed University of Artificial Intelligence, UAE,
[4]University of Padova, Italy, [5]University of Trento, Italy
{mhasanain, fialam, hmubarak, wzaghouani}@hbku.edu.qa,
preslav.nakov@mbzuai.ac.ae, dasan@math.unipd.it, abdel.fraihat@gmail.com

Abstract

We present an overview of the ArAIEval shared task, organized as part of the first ArabicNLP 2023 conference co-located with EMNLP 2023. ArAIEval offers two tasks over Arabic text: (*i*) persuasion technique detection, focusing on identifying persuasion techniques in tweets and news articles, and (*ii*) disinformation detection in binary and multiclass setups over tweets. A total of 20 teams participated in the final evaluation phase, with 14 and 16 teams participating in Tasks 1 and 2, respectively. Across both tasks, we observed that fine-tuning transformer models such as AraBERT was at the core of the majority of the participating systems. We provide a description of the task setup, including a description of the dataset construction and the evaluation setup. We further give a brief overview of the participating systems. All datasets and evaluation scripts from the shared task are released to the research community.[1] We hope this will enable further research on these important tasks in Arabic.

1 Introduction

Social media has become one of the predominant communication channels for freely sharing content online. With this freedom, misuse has emerged, turning social media platforms into potential grounds for sharing inappropriate posts, misinformation, and disinformation (Zhou et al., 2016; Alam et al., 2022a; Sharma et al., 2022). Malicious users can disseminate disinformative content, such as hate-speech, rumors, and spam, to gain social and political agendas or to harm individuals, entities and organizations. Such content can inflame tension between different groups and ignite violence among their members, making early detection and prevention essential.

Previous successful attempts to address such kinds of problems at a large scale over Arabic content include offensive and hate speech detection shared tasks (Zampieri et al., 2020; Mubarak et al., 2020b).

Social media content designed to promote hidden agendas is not limited to disinformation. In the past years, propaganda has been widely used as well, to influence and/or mislead the audience, which became a major concern for different stakeholders, social media platforms and government agencies. News reporting in the mainstream media also exhibits a similar phenomenon, where a variety of persuasion techniques (Miller, 1939) are used to promote a particular editorial agenda. To address this problem, the research area of "computational propaganda" has emerged aimed at automatically identify such techniques in textual, visual and multimodal (e.g., memes) content. Da San Martino et al. (2019) curated a set of persuasion techniques, such as *Loaded Language*, *Appeal to Fear*, *Straw Man* and *Red Herring*. The focus of the work was mainly on textual content (i.e., newspaper articles). Following this prior work, in 2021, Dimitrov et al. (2021) organized a shared task on propaganda techniques in memes. These efforts mainly focused on English. To enrich the Arabic AI research, we have organized a shared task on detection of fine-grained propaganda techniques for Arabic, which attracted many participants (Alam et al., 2022b).

Following the success of our previous shared tasks (Alam et al., 2022b; Zampieri et al., 2020; Mubarak et al., 2020b), and given the great interest from the community in further pushing research in this domain, this year we organize the **Ar**abic **AI Eval**uation (**ArAIEval**) shared task covering the following two tasks: *(i)* persuasion technique detection over tweets and news articles, and *(ii)* disinformation detection over tweets.

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

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 483–493

This edition of the shared task has attracted wide participation. The task was run in two phases: *(i)* the development phase with 38 registrations, and 14 teams submitting their systems; and *(ii)* the evaluation phase with 25 registrations, and 20 teams submitting their systems. In the remainder of this paper, we define each of the two tasks, describe the Arabic evaluation datasets that were manually constructed, and provide overview of participating systems and their official scores.

2 Related Work

2.1 Persuasion Techniques Detection

The history of studying propaganda can be traced back to the 17th century, where the focus was to understand whether manipulation techniques were used during public events at theaters, festivals, and games (Margolin, 1979; Casey, 1994). Since then, the study of propaganda has spanned across various disciplines including history, journalism, political science, sociology, and psychology (Jowett and O'donnell, 2018). Different disciplines explored propaganda for varied purposes; for instance, in political science, it is studied to analyze the ideologies of practitioners and to understand the impact of information dissemination on public opinion.

Over the last few decades, the current information ecosystem has undergone significant changes due to the emergence of social media platforms, which have become breeding grounds for the creation and dissemination of misinformation and propaganda. Consequently, there has been research aimed at understanding and automatically detecting such content by defining the rhetorical and psychological techniques employed on online platforms.

Most computational approaches for automatic detection involve identifying whether textual content contains propaganda (Barrón-Cedeno et al., 2019), identifying propagandistic techniques (Habernal et al., 2017, 2018), and detecting propagandistic text spans in news articles (Da San Martino et al., 2019, 2020). The majority of these studies have primarily focused on English. To address this issue in multilingual settings, a shared task was recently organized, focusing on nine languages (Piskorski et al., 2023). The outcomes of such initiatives highlight the importance of multilingual models. For instance, Hasanain et al. (2023) show that multilingual models significantly outperform monolingual models, even for languages unseen during training.

Other relevant shared tasks include those focusing on multimodality. Dimitrov et al. (2021) organized SemEval-2021 Task 6 on the propaganda detection in memes, which comprises a multimodal setup involving both text and images.

Along such initiatives, we have primarily focused on Arabic content. The propaganda shared task, co-located with WANLP 2022, was mainly focused on tweets in both binary and multilabel settings (Alam et al., 2022b). This year, we have expanded it on a larger scale with a larger dataset, focusing on news articles and tweets.

2.2 Disinformation Detection

Disinformation is relatively a new term and it is defined as "*fabricated or deliberately manipulated text/speech/visual context, and also intentionally created conspiracy theories or rumors*" (Ireton and Posetti, 2018). There have been several studies on the automatic detection of bad content on social media, including hate speech (Fortuna and Nunes, 2018), harmful content (Alam et al., 2021, 2022a), rumors (Meel and Vishwakarma, 2020), and offensive language (Husain and Uzuner, 2021).

In the context of Arabic social media, numerous researchers have employed different approaches to disinformation detection. For instance, Boulouard et al. (2022) investigated disinformation detection, particularly hate-speech and offensive content detection, on Arabic social media.

For this shared task on disinformation detection, our work is inspired by Mubarak et al. (2023), which primarily focused on detecting disinformative tweets that are most likely to be deleted.

3 Task 1: Propaganda Detection

The goal of this task is to identify the persuasion techniques present in a piece of text. It targets multi-genre content, including tweets and paragraphs from news articles, as persuasion techniques are commonly used within these domains. The task is organized into two subtasks.

3.1 Subtasks

Subtask 1A: Given a text snippet, identify whether it contains content with any persuasion technique. This is a *binary classification* task.

Subtask 1B: Given a text snippet, identify the propaganda techniques used in it. This is a *multilabel classification* task.

	Train	Dev	Test
true	1918 (79%)	202 (78%)	331 (66%)
false	509 (21%)	57 (22%)	172 (34%)
Total	**2427**	**259**	**503**

Table 1: Distribution of Subtask **1A** dataset. In parentheses, we show the percentage of a label in a split.

3.2 Dataset

To construct the annotated dataset for this task, we collected different datasets consisting of tweets and news articles, as discussed below.

Tweets: We start from the same tweets dataset collected from Twitter accounts of Arabic news sources, as described in the previous edition of the shared task (Alam et al., 2022b). We randomly sampled a subset of 156 tweets for annotation to construct the *testing subset* of this task. The number of tweets selected for annotation was decided based on time and cost required for annotation.

News paragraphs: We select news articles from an existing dataset, AraFacts (Ali et al., 2021), that contains claims verified by Arabic fact-checking websites, and each claim is associated with web pages propagating or negating the claim. We keep the pages that are from news domains in the set (e.g., www.alquds.co.uk). We automatically parsed these news articles and split them into paragraphs based on blank lines.

Data annotation: For both tweets and paragraphs, we follow the same annotation process to identify the persuasion techniques in a snippet. The process includes two phases: (*i*) three annotators independently annotated the same text snippet, through an annotation interface designed for the task, and (*ii*) two consolidators reviewed the annotations and produced the gold annotations. Annotators were recruited and trained for the task in-house. We annotate text by a set of 23 persuasion techniques that is adopted from existing research (Piskorski et al., 2023). We should note here that multiple techniques can be found in the same text snippet. *For Subtask 1A (binary classification)*, the labels were generated by assigning a positive label (true) to every text snippet that had at least one persuasion technique, and a negative label was given otherwise. Below we give an example subset of the persuasion techniques, and briefly summarize them:

1. **Loaded language:** using specific emotionally-loaded words or phrases (positive or negative) to convince the audience that an argument is valid.
2. **Appeal to Fear, Prejudice:** building support or rejection for an idea by instilling fear or repulsion towards it, or to an alternative idea.
3. **Strawman:** giving the impression that an argument is being refuted, whereas the real subject of the argument was not addressed or refuted, but instead was replaced with a different one.

Persuasion Technique	Train (2427)	Dev (259)	Test (503)
Loaded Language	1574	176	253
Name Calling or Labelling	692	77	133
Questioning the Reputation	383	43	89
Exaggeration or Minimisation	292	33	40
Obfuscation, Intentional Vagueness, Confusion	240	28	25
Casting Doubt	143	16	21
Causal Oversimplification	128	15	12
Appeal to Fear, Prejudice	108	12	15
Slogans	70	8	25
Flag Waving	63	7	25
Appeal to Hypocrisy	56	7	17
Appeal to Values	37	4	29
Appeal to Authority	48	5	14
False Dilemma or No Choice	32	3	6
Consequential Oversimplification	33	3	3
Conversation Killer	28	3	7
Repetition	25	3	6
Guilt by Association	13	1	1
Appeal to Time	10	2	2
Whataboutism	9	1	2
Red Herring	8	1	3
Strawman	6	1	2
Appeal to Popularity	2	1	1
No Technique	*509*	*57*	*172*
Total	**4509**	**507**	**903**

Table 2: Distribution of the techniques for the Subtask **1B** dataset: sorted by total frequency over all splits. In parentheses, we show the total number of documents in a split.

Data splits: The full set of annotated paragraphs is divided into three subsets: train, development, and test, using a stratified splitting approach to ensure that the distribution of persuasion techniques is consistent across the splits. For the tweets set, we split the full annotated tweet set from the previous edition of the lab (Alam et al., 2022b) into train and development subsets, while the test set is annotated for this shared task. Finally, we construct the multi-genre subsets for the task by merging the sets of paragraphs and tweets.

Statistics: In Tables 1 and 2 we show the distribution of labels across splits for Task 1.

Team		Subtask		Model								Misc.			
		1A	1B	AraBERT	MARBERT	ArabicBERT	BERT	RoBERTa	XLM-RoBERTa	AraELECTRA	GPT	Data augm.	Preprocessing	Ensamble	Loss Funct.
HTE	(Khaldi and Bouklouha, 2023)	1	5	✓	✓										✓
KnowTellConvince	(Veeramani et al., 2023)	2				✓						✓		✓	✓
rematchka	(Abdel-Salam, 2023)	3	2	✓	✓										✓
UL & UM6P	(Lamsiyah et al., 2023)	4	1	✓	✓										✓
Itri Amigos	(Ahmed et al., 2023)	5	4	✓											
Raphael	(Utsav et al., 2023)	6	6		✓			✓			✓				
Frank	(Azizov, 2023)	7			✓		✓	✓					✓		
Mavericks	(Mangalvedhekar et al., 2023)	8		✓						✓			✓	✓	
Nexus	(Xiao and Alam, 2023)	9		✓	✓								✓		
AAST-NLP	(ElSayed et al., 2023)	11	3	✓	✓							✓	✓		✓
ReDASPersuasion	(Qachfar and Verma, 2023)	13	7						✓				✓		
Legend	(Ojo et al., 2023)	14							✓						

Table 3: Overview of the systems for **Task 1**. Numbers under the subtask code indicate the position of the team in the official ranking. Data augm.: Data augmentation. Loss Funct.: Experiments with a variety of loss functions.

	Team	Micro F1	Macro F1		Team	Micro F1	Macro F1
	Subtask 1A				**Subtask 1B**		
1	HTE	0.7634	0.7321	1	UL & UM6P	0.5666	0.2156
2	KnowTellConvince	0.7575	0.7282	2	rematchka	0.5658	0.2497
3	rematchka	0.7555	0.7309	3	AAST-NLP	0.5522	0.1425
4	UL & UM6P	0.7515	0.7186	4	Itri Amigos	0.5506	0.1839
5	Itri Amigos	0.7495	0.7225	5	HTE	0.5412	0.0979
6	Raphael	0.7475	0.7221	6	Raphael	0.5347	0.1772
7	Frank	0.7455	0.7173	7	ReDASPersuasion	0.4523	0.0568
8	Mavericks	0.7416	0.7031	8	*Baseline (Majority)*	0.3599	0.0279
9	Nexus	0.7396	0.6929	9	*Baseline (Random)*	0.0868	0.0584
10	superMario	0.7316	0.7098	10	pakapro	0.0854	0.0563
11	AAST-NLP	0.7237	0.6693				
12	*Baseline (Majority)*	0.6581	0.3969				
13	ReDASPersuasion	0.6581	0.3969				
14	Legend	0.6402	0.4647				
15	pakapro	0.5030	0.4940				
16	*Baseline (Random)*	0.4771	0.4598				

Table 4: Official results for **Task 1**. Runs ranked by the official measure: Micro F1.

3.3 Evaluation Setup

The task was organized into two phases:

- **Development phase**: we released the train and development subsets, and participants submitted runs on the development set through a competition on Codalab [2].
- **Test phase**: we released the official test subset, and the participants were given a few days to submit their final predictions through a competition on Codalab.[3] Only the latest submission from each team was considered official and was used for the final team ranking.

[2] https://codalab.lisn.upsaclay.fr/competitions/14563

[3] https://codalab.lisn.upsaclay.fr/competitions/15099

Measures: We measure the performance of the participating systems, for all subtasks, using micro-averaged F1 as the official evaluation measure of the shared task, as these are multiclass/multilabel problems, where the labels are imbalanced. We also report macro-averaged F1, as an unofficial evaluation measure.

3.4 Overview of Participating Systems and Results

A total of 14 and 8 teams submitted runs for Subtask 1A and 1B, respectively, with 8 teams making submissions for both subtasks. Table 3, overviews 12 of the participating systems for which a description paper was submitted. Table 4 presents the results and rankings of *all* systems.

	Train	Dev	Test
Disinfo	2656 (19%)	397 (19%)	876 (23%)
Not-disinfo	11491 (81%)	1718 (81%)	2853 (77%)
Total	**14147**	**2115**	**3729**

Table 5: Distribution of Subtask **2A** dataset. In parentheses, we show the percentage of a label in a split.

Fine-tuning pre-trained Arabic models (specifically AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021)) was the most common system architecture. However, we observed that several systems also experimented with a variety of loss functions for model training to handle characteristics of the training dataset, like label imbalance (Lamsiyah et al., 2023; Khaldi and Bouklouha, 2023; Veeramani et al., 2023; Abdel-Salam, 2023; ElSayed et al., 2023).

When comparing the performance to the previous edition (Alam et al., 2022b) for the multilabel subtask, we observe that this year's Subtask 1B is much more challenging. In the previous edition, the best system achieved a Micro F1 of 0.649, whereas this year it is 0.566, keeping in mind that the dataset is different and may not be exactly comparable.

4 Task 2: Disinformation Detection

This task targeted tweets and was organized into two subtasks, as discussed below.

4.1 Subtasks

Subtask 2A: Given a tweet, identify whether it is disinformative. This is a *binary classification* task.

Subtask 2B: Given a tweet, detect the fine-grained disinformation class, if any. This is a *multiclass classification* task. The fine-grained labels include *hate-speech*, *offensive*, *rumor*, and *spam*.

4.2 Dataset

We have constructed an annotated dataset composed of $20K$ tweets, labeled as disinformative or not-disinformative, along with fine-grained categories for the disinformative set. These tweets are related to COVID-19 and were collected in February and March 2020. We followed the annotation guidelines described in (Mubarak et al., 2020b), (Zampieri et al., 2020), (Mubarak et al., 2022), and (Mubarak et al., 2020a), for hate speech, offensive content, rumor, and spam classes, respectively. More details about data collection and annotation can be found in (Mubarak et al., 2023). Tables 5 and 6 display the statistics of the dataset.

	Train	Dev	Test
HS	1512 (57%)	226 (57%)	442 (50%)
Off	500 (19%)	75 (19%)	160 (18%)
Rumor	191 (7%)	28 (7%)	33 (4%)
Spam	453 (17%)	68 (17%)	241 (28%)
Total	**2656**	**397**	**876**

Table 6: Distribution of Subtask **2B** dataset. In parentheses, we show the percentage of a label in a split.

4.3 Evaluation Setup and Measures

Similar to Task 1, we also conducted this task in two phases as discussed in Section 3.3. Systems were valuated using Micro F1 as the official measure, while also reporting Macro F1.

4.4 Overview of Participating Systems and Results

Table 7 and 8 overviews the submitted systems, and the official results and ranking, respectively. A total of 15 and 11 teams participated in Subtask 2A and 2B, respectively, out of which, 10 made submissions for both subtasks. Out of 17 teams, 13 outperformed the majority baseline for Subtask 2A, whereas out of 11 teams, 9 outperformed the majority baseline for Subtask 2B. These subtasks were dominated by transformer models as observed in Table 7. The most commonly used model was AraBERT (Antoun et al., 2020), followed by MARBERT (Abdul-Mageed et al., 2021), ARBERT(Abdul-Mageed et al., 2021), and QARiB (Abdelali et al., 2021). Half of the participants employed preprocessing techniques, and the top-performing teams utilized data augmentation.

5 Participating Systems

AAST-NLP (ElSayed et al., 2023) The team experimented with several transformer-based models, including MARBERT (Abdul-Mageed et al., 2021), ARBERT (Abdul-Mageed et al., 2021), and AraBERT (Antoun et al., 2020). AraBERT outperformed the others across all subtasks. Preprocessing was applied using the AraBERT preprocessor. Tweet tags, emojis, and Arabic stopwords were removed. For the final submission, binary cross entropy was selected for multilabel classification (Subtask 1B), while Dice loss was chosen for the remaining three subtasks. Although the team tried data augmentation with contextual word embeddings and a hybrid approach combining AraBERT with a CNN-BILSTM, these did not improve accuracy.

Team		Subtask		Model												Misc.	
		2A	2B	AraBERT	MARBERT	ARBERT	QARiB	CAMeLBERT	BERT	RoBERTa	XLM-RoBERTa	DistilBERT	AraELECTRA	LSTM	SVM	Data augm.	Preprocessing
DetectiveRedasers	(Tuck et al., 2023)	1	1	✓	✓		✓	✓								✓	✓
AAST-NLP	(ElSayed et al., 2023)	2	3	✓	✓	✓								✓		✓	✓
UL & UM6P	(Lamsiyah et al., 2023)	3	2	✓	✓	✓											
rematchka	(Abdel-Salam, 2023)	4	4	✓	✓	✓										✓	
PD-AR	(Deka and Revi, 2023)	5	6	✓				✓	✓	✓	✓						✓
Mavericks	(Mangalvedhekar et al., 2023)	7		✓									✓				✓
Itri Amigos	(Ahmed et al., 2023)	8	7	✓													✓
KnowTellConvince	(Veeramani et al., 2023)	9	8	✓													
Nexus	(Xiao and Alam, 2023)	10		✓	✓		✓										
PTUK-HULAT	(Jaber and Martinez, 2023)	11							✓			✓					✓
Frank	(Azizov, 2023)	12			✓				✓	✓							
USTHB	(Mohamed et al., 2023)	13	9												✓		
AraDetector	(Ahmed Bahaaulddin A. et al., 2023)	15		✓	✓		✓										✓

Table 7: Overview of the systems for **Task 2**. The numbers under the subtask code indicate the position of the team in the official ranking. Data augm.: Data augmentation.

	Team	Micro F1	Macro F1		Team	Micro F1	Macro F1
	Subtask 2A				**Subtask 2B**		
1	DetectiveRedasers	0.9048	0.8626	1	DetectiveRedasers	0.8356	0.7541
2	AAST-NLP	0.9043	0.8634	2	UL & UM6P	0.8333	0.7388
3	UL & UM6P	0.9040	0.8645	3	AAST-NLP	0.8253	0.7283
4	rematchka	0.9040	0.8614	4	rematchka	0.8219	0.7156
5	PD-AR	0.9021	0.8595	5	superMario	0.8208	0.7031
6	superMario	0.9019	0.8625	6	PD-AR	0.8174	0.7209
7	Mavericks	0.9010	0.8606	7	Itri Amigos	0.8139	0.7220
8	Itri Amigos	0.8984	0.8468	8	KnowTellConvince	0.8071	0.6888
9	KnowTellConvince	0.8938	0.8460	9	USTHB	0.5046	0.1677
10	Nexus	0.8935	0.8459	10	*Baseline (Majority)*	0.5046	0.1677
11	PTUK-HULAT	0.8675	0.7992	11	Ankit	0.4167	0.1993
12	Frank	0.8163	0.6378	12	*Baseline (Random)*	0.2603	0.2243
13	USTHB	0.7670	0.4418	13	pakapro	0.2317	0.1978
14	*Baseline (Majority)*	0.7651	0.4335				
15	AraDetector	0.7487	0.6498				
16	*Baseline (Random)*	0.5154	0.4764				
17	pakapro	0.4996	0.4596				

Table 8: Official results for **Task 2**. Runs ranked by the official measure: Micro F1.

AraDetector (Ahmed Bahaaulddin A. et al., 2023) The team tackled Subtask 2A using an ensemble of three classifiers: MARBERT model fine-tuned on the training data, and GPT-4 (OpenAI, 2023) in zero-shot and few-shot settings. A majority voting approach was then used to merge the binary predictions of the three classifiers. The results on the development set showed that GPT-4 in zero-shot setting outperforms the ensemble model by the Micro F1 measure.

DetectiveRedasers (Tuck et al., 2023) The team participated in subtasks 2A and 2B following a two-fold methodology. First, they conducted comprehensive preprocessing, addressing challenges like code-switching and use of emoji in tweets. Non-Arabic portions of the tweets were then automatically translated into Arabic. Instead of removing emojis and hashtags, these were converted into Arabic descriptive text to preserve the sentiment of the tweets. For Subtask 2A, the team used AraBERT-Covid19[4] with hyperparameters optimized through the optimization framework Optuna. As for Subtask 2B, a soft voting ensemble method is used with five optimized AraBERTv02-Twitter (Antoun et al., 2020) models, each with identical hyperparameters and architecture, only differing by random initialization. AraBERTv02-Twitter was selected since it is based on the effective AraBERT mode, with continued pre-training on $60M$ Arabic tweets, making it suitable for Subtask 2B focused on tweets.

[4] https://huggingface.co/moha/arabert_arabic_covid19

Frank (Azizov, 2023) After preprocessing using AraBERT preprocessor, multilingual BERT (Devlin et al., 2018) was fine-tuned for Subtask 1A, and MARBERT was fine-tuned for Subtask 2A.

HTE (Khaldi and Bouklouha, 2023) Participating in Subtask 1A, the team fine-tuned the MARBERT model in a multitask setting: a primary binary classification task to identify the presence of persuasive techniques in text generally, and an auxiliary task focused on classifying texts based on their type (tweet or news). It was expected that the auxiliary task would help the primary task in learning specific lexical and syntactic information about tweets or news related to persuasive content. Given the imbalance in the dataset, the team employed focal loss to optimize both tasks. On the test set, the system ranked highest on the leaderboard.

Itri Amigos (Ahmed et al., 2023) The team submitted runs for all four subtasks. Preprocessing was applied using AraBERT preprocessor. Further preprocessing was done for all subtasks but 1B, where links and mentions were removed. For subtasks 1A and 1B, the team fine-tuned the AraBERTv2 transformer model. To address the class imbalance in the datasets, class weights were incorporated during training. As for subtasks 2A and 2B that are mainly targeting tweets, AraBERTv02-Twitter was fine-tuned for the tasks.

KnowTellConvince (Veeramani et al., 2023) The team participated in subtasks 1A, 2A and 2B using an ensemble of the following four models. *(i)* fine-tuned BERT Arabic base model (Safaya et al., 2020) with a contrastive loss function; *(ii)* fine-tuned BERT Arabic base model with a cross entropy loss function; *(iii)* fine-tuned BERT Arabic base on XNLI dataset to capture nuances relevant to sentiment as part of the system architecture; and *(iv)* a model utilizing sentence embeddings from BERT Arabic base followed by computing cosine similarity between pairs of sentences from the data, that finally goes through Gaussian Error Linear Unit (GELU) activation.

Legend (Ojo et al., 2023) team participated in Task 1, in which XLM-RoBERTa was implemented. To address the class imbalance in the dataset, the team adjusted the learning process using class weights. A learning rate scheduler was implemented to dynamically adjust the learning rate during training. Specifically, they used a StepLR scheduler with a reduction factor of 0.85 applied every 2 epochs. This scheduling strategy contributes to the training stability and the controlled convergence.

Mavericks (Mangalvedhekar et al., 2023) Targeting subtasks 1A and 2A, several transformer-based models were fine-tuned on the provided dataset. The models include: AraBERT, MARBERT and AraELECTRA (Antoun et al., 2021). Ensembling was utilized using hard voting, where the majority vote of all the predictions is selected as the final prediction.

Nexus (Xiao and Alam, 2023) The team explored performance of fine-tuning several pre-trained language models (PLMs) including AraBERT, MARBERT, and QARiB in subtasks 1A and 2A. In addition to that, experiments with GPT-4 (OpenAI, 2023) in both zero-shot and few-shot settings were conducted for both subtasks. Performance of the GPT-4 model was notably lower than the fine-tuned models.

PD-AR (Deka and Revi, 2023) For both sub-tasks 2A and 2B, the team employed the AraBERTv0.2-Twitter-base model and utilized the provided training and development sets to train the model. Before training, some preprocessing of the text was performed. Compared to fine-tuning several other PLMs such as XLM-RoBERTa (Conneau et al., 2020), the Arabic-specific model showed significantly improved performance.

PTUK-HULAT (Jaber and Martinez, 2023) The team participated in Subtask 2A, in which they fine-tuned a multilingual DistilBERT model on the corresponding binary classification data. They then used the fine-tuned model to predict whether a tweet is dis-informative or not.

Raphael (Utsav et al., 2023) For both subtasks 1A and 1B, they used MARBERT as the encoder. In addition to that, they used GPT-3.5 (Brown et al., 2020) in order to generate English descriptions of the Arabic texts and to provide tone and emotional analysis. The resulting English text and tone descriptions were then encoded using RoBERTa (Liu et al., 2019) and were further concatenated to the MARBERT encodings. Finally, the full embeddings were passed to a binary classification head and to multilabel classification heads for Subtasks 1A and 1B, respectively.

ReDASPersuasion (Qachfar and Verma, 2023) The initial structure of the system has three main components: *(i)* A multilingual transformer model that tokenizes the input and produced a [CLS] embedding output; *(ii)* A feature engineering module designed to extract language-agnostic features for persuasion detection; *(iii)* A multi-label classification head that integrates the first and the second components, using a sigmoid activation and cross entropy loss. For subtasks 1A and 1B, the system was paired with DistilBERT (Sanh et al., 2019) for the official submission, but follow-up experiments for Subtask 1A showed that using XLM-RoBERTa, yielded the best Micro F1 score on test.

rematchka (Abdel-Salam, 2023) For all subtasks, ARBERTv2 (Abdul-Mageed et al., 2021), AraBERTv2, and MARBERT models were trained on the provided datasets. For Subtask 1A, different techniques such as fast gradient methods and contrastive learning were applied. Moreover, the team employed back-translation between Arabic and English for data augmentation. As for Subtask 1B, different loss functions, including Asymmetric loss and Distribution Balanced loss were tested. Moreover, a balanced data-sampler for multilabel datasets was used. Fro both subtasks, prefix tuning was used for model training.

UL & UM6P (Lamsiyah et al., 2023) used an Arabic pre-trained transformer combined with a classifier. The performance of three transformer models was evaluated for sentence encoding. For Subtask 1A, the MARBERTv2 encoder was used, and the model was trained with cross-entropy and regularized Mixup (RegMixup) loss functions. For Subtask 1B, the AraBERT-Twitter-v2 encoder was used, and the model was trained with the asymmetric multi-label loss. The significant impact of the training objective and text encoder on the model's performance was highlighted by the results. For Subtask 2A, the AraBERT-Twitter-v2 encoder was used, and the model was trained with cross-entropy loss. For Subtask 2B, the MARBERTv2 encoder was used, and the model was trained with the Focal Tversky loss.

USTHB (Mohamed et al., 2023) For both subtasks 2A and 2B, the system start with extensive preprocessing of the data. Then, the FastText model is used for feature extraction in addition to TF-IDF to vectorize the data. SVM was then trained as a classifier.

6 Conclusion and Future Work

We presented an overview of the ArAIEval shared task at the ArabicNLP 2023 conference, targeting two shared tasks: *(i)* persuasion technique detection, and *(ii)* disinformation detection. The task attracted the attention of many teams: a total of 25 teams registered to participate during the evaluation phase, with 14 and 16 teams eventually making an official submission on the test set for tasks 1 and 2, respectively. Finally, 17 teams submitted a task description paper. Task 1 aimed to identify the propaganda techniques used in multi-genre text snippets, including tweets and news articles, in both binary and multilabel settings. On the other hand, Task 2 aimed to detect disinformation in tweets in both binary and multiclass settings. For both tasks, the majority of the systems fine-tuned pre-trained Arabic language models and used standard pre-processing. Several systems explored different loss functions, while a handful of systems utilized data augmentation and ensemble methods.

Given the success of the task this year, we plan to run a future edition with an increased data size, and with wider coverage of domains, countries, and Arabic dialects. We are also considering implementing a multi-granularity persuasion techniques detection setting.

Limitations

Task 1 was limited to binary an multilabel classification. A natural next step would have been to also run a span detection subtask, which is a more complex task. This was left for future editions of ArAIEval. This is to ensure enough participation after building a strong community working on propaganda detection over Arabic content in the less complex setups. As for Task 2, we observe the systems achieved significantly high performance, even in the more challenging multiclass setup. One potential reason might be that the dataset developed was too easy. Investigating how to make this task more challenging while reflecting real-world scenarios was not in this edition of the shared task, but is within our future plan.

Acknowledgments

M. Hasanain's contribution was funded by NPRP grant 14C-0916-210015, and W. Zaghouani contribution was partially funded by NPRP grant 13S-0206-200281 Both projects are funded by the Qatar

National Research Fund (a member of Qatar Foundation).

Part of this work was also funded by Qatar Foundation's IDKT Fund TDF 03-1209-210013: *Tanbih: Get to Know What You Are Reading*.

We thank Fatema Akter and Hussein Mohsin Al-Dobashi for helping with the persuasion techniques annotations.

The findings herein are solely the responsibility of the authors.

References

Reem Abdel-Salam. 2023. rematchka at ArAIEval Shared Task: Prefix-Tuning & Prompt-tuning for Improved Detection of Propaganda and Disinformation in Arabic Social Media Content. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

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

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.

Nouman Ahmed, Natalia Flechas Manrique, and Jehad Oumer. 2023. Itri Amigos at ArAIEval Shared Task: Transformer vs. Compression-Based Models for Persuasion Techniques and Disinformation Detection. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Wahhab Ahmed Bahaaulddin A., Sabeeh Vian, Belhaj Hanan Mohamed, Sibaee Serry, Samar Ahmad, Khurfan Ibrahim, and Alharbi Abdullah I. 2023. AraDetector at ArAIEval Shared Task: An Ensemble of Arabic-specific pre-trained BERT and GPT-4 for Arabic Disinformation Detection . In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Preslav Nakov, and Giovanni Da San Martino. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, WANLP '22, Abu Dhabi, UAE.

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.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: the first large arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Dilshod Azizov. 2023. Frank at ArAIEval Shared Task: Arabic Disinformation and Persuasion: Power of Pretrained Models. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Zakaria Boulouard, Mariya Ouaissa, Mariyam Ouaissa, Moez Krichen, Mutiq Almutiq, and Gasmi Karim. 2022. Detecting hateful and offensive speech in arabic social media using transfer learning. *Applied Sciences*, 12:12823.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ralph D. Casey. 1994. What is propaganda? *historians.org*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Giovanni Da San Martino, Alberto Barrón-Cedeno, 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.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China.

Pritam Deka and Ashwathy Revi. 2023. PD-AR at ArAIEval Shared Task: Persuasion techniques detection: an interdisciplinary approach to identifying and counteracting manipulative strategies. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. 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*.

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.

Ahmed ElSayed, Omar Nasr, and Nour Eldin Elmadany. 2023. AAST-NLP at ArAIEval Shared Task: Tackling Persuasion Technique and Disinformation Detection using Pre-Trained Language Models On Imbalanced Datasets. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3329–3335.

Maram Hasanain, Ahmed El-Shangiti, Rabindra Nath Nandi, Preslav Nakov, and Firoj Alam. 2023. QCRI at SemEval-2023 task 3: News genre, framing and persuasion techniques detection using multilingual models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1237–1244, Toronto, Canada. Association for Computational Linguistics.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.

Areej Jaber and Paloma Martinez. 2023. PTUK-HULAT at ArAIEval Shared Task: Fine-tuned Distilbert to Predict Disinformative Tweets. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Garth S Jowett and Victoria O'donnell. 2018. *Propaganda & persuasion*. Sage publications.

Hadjer Khaldi and Taqiy Eddine Bouklouha. 2023. HTE at ArAIEval Shared Task: Persuasion techniques detection: an interdisciplinary approach to identifying and counteracting manipulative strategies. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Salima Lamsiyah, El Mahdaouy Abdelkader, Hamza Alami, Ismail Berrada, and Christoph Schommer. 2023. UL& UM6P at ArAIEval Shared Task: Transformer-based model for Persuasion Techniques and Disinformation detection in Arabic. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. 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*.

Sudeep Mangalvedhekar, Kshitij Deshpande, Yash Patwardhan, Vedant Deshpande, and Ravindra Murumkar. 2023. Mavericks at ArAIEval Shared Task: Towards a Safer Digital Space - Transformer Ensemble Models Tackling Deception and Persuasion. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

V. Margolin. 1979. The visual rhetoric of propaganda. *Information Design Journal*, 1:107–122.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-

arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Clyde R. Miller. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning.

Lichouri Mohamed, Lounnas Khaled, Zitouni Aicha, Latrache Houda, and Djeradi Rachida. 2023. USTHB at ArAIEval Shared Task: Disinformation Detection System based on Linguistic Feature Concatenation. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

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*.

Hamdy Mubarak, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish. 2020a. Spam detection on arabic twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 237–251. Springer.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020b. Overview of osact4 arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pages 48–52.

Hamdy Mubarak, Sabit Hassan, Shammur Absar Chowdhury, and Firoj Alam. 2022. ArCovidVac: Analyzing Arabic tweets about COVID-19 vaccination. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3220–3230, Marseille, France. European Language Resources Association.

Olumide E. Ojo, Olaronke O. Adebanji, Hiram Calvo, Damian O. Dieke, Olumuyiwa E. Ojo, Seye E. Akinsanya, Tolulope O. Abiola, and Anna Feldman. 2023. Legend at ArAIEval Shared Task: Persuasion Technique Detection using a Language-Agnostic Text Representation Model. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Fatima Zahra Qachfar and Rakesh M. Verma. 2023. ReDASPersuasion at ArAIEval Shared Task: Multilingual and Monolingual Models For Arabic Persuasion Detection. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI '22, pages 5597–5606, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Bryan E. Tuck, Fatima Zahra Qachfar, Dainis Boumber, and Rakesh M. Verma. 2023. DetectiveRedasers at ArAIEval Shared Task: Leveraging Transformer Ensembles for Arabic Deception Detection. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Shukla Utsav, Tiwari Shailendra, and Vyas Manan. 2023. Raphael at ArAIEval Shared Task: Understanding Persuasive Language and Tone, an LLM Approach. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. KnowTellConvince at ArAIEval Shared Task: Disinformation and Persuasion Detection using Similar and Contrastive Representation Alignment. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Yunze Xiao and Firoj Alam. 2023. Nexus at ArAIEval Shared Task: Fine-Tuning Arabic Language Models for Propaganda and Disinformation Detection. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Lu Zhou, Wenbo Wang, and Keke Chen. 2016. Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 603–612, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.