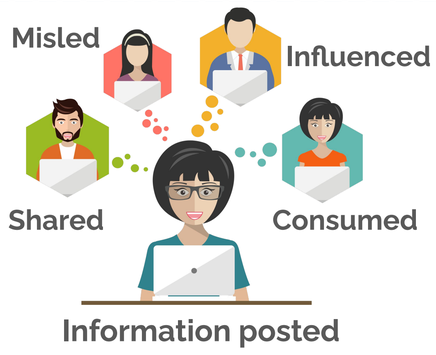
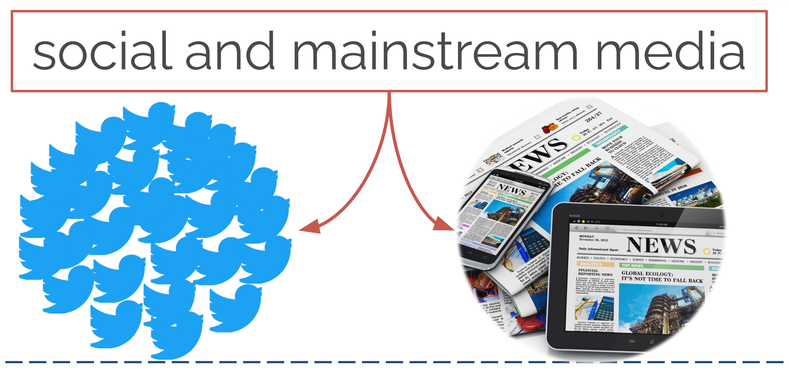
ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, Abed Alhakim Freihat

https://araieval.gitlab.io

- Task 1: Persuasion Technique Detection
- Task 2: Disinformation Detection





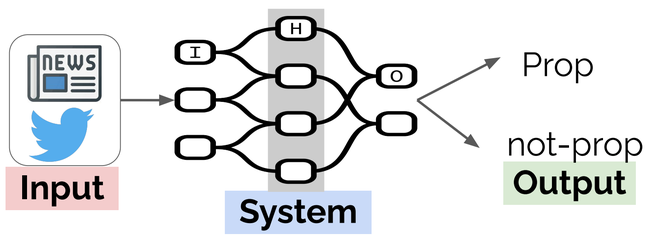
Tasks

1 Persuasion Technique Detection

Communication that **deliberately** misrepresent symbols, appealing to emotions and prejudices and bypassing rational thought, to **influence** its audience **towards a specific goal**

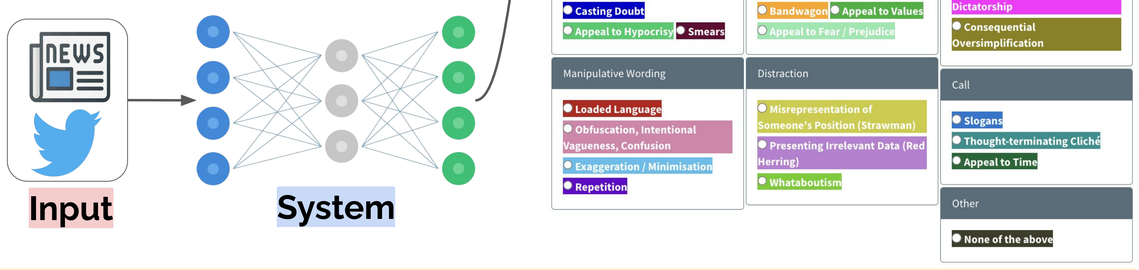
Binary Classification Task

Subtask 1A: Given a multigenre (tweet and news paragraphs of the news articles) snippet, identify whether it contains content with persuasion technique. This is a **binary classification task.**



Multilabel Classification Task

Subtask 1B: Given a multigenre (tweet and news paragraphs of the news articles) snippet, identify the propaganda techniques used in it. This is a **multilabel classification task.**



Data Collection:

- **Tweets** collected from different accounts of Arabic news sources (Alam et al., 2022b)
- **News paragraphs** selected from news articles (Hasanain et al. 2023)
  - **a.** AraFacts (Ali et al., 2021) **b.** in-house news articles collection

Annotation:

- **Phase 1:** Individual annotators annotate the dataset
- **Phase 2:** Consolidation is done with expert annotators to resolve the disagreement and ensure quality

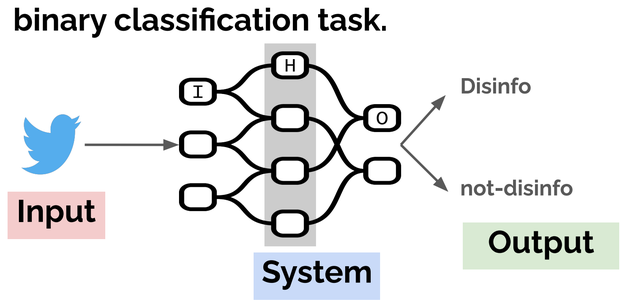
Dataset: Subtask 1A (3,189); Subtask 1B (5,919)

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"*

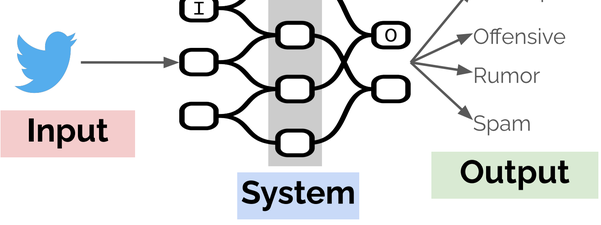
Binary Classification Task

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



Multiclass 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.



Data Collection:

- **Arabic tweets** collected in February & March 2020 using keyword Corona
- Selected tweets that were deleted after posting
- Manually annotated **22K** deleted and non-deleted tweets with different categories

Dataset:

- Subtask 2A: ~20K
- Subtask 2B: ~4K

Class	Example
HS*	أنا مؤمن تماماً أن الصينيين سبب تفشي أمراض مثل سارس و كورونا I strongly believe that the Chinese caused the outbreak of diseases such as SARS and Corona
Off*	لسانها اوصخ من كورونا Her tongue is dirtier than Corona
Rumor	دواء الملاريا هو الذي يعالج كورونا بنسبة 001% Malaria medicine cures Corona with 100% efficiency

Results (Task 1)

	Team	Micro F1	Macro F1
	Subtask 1A		
1	HTE	0.7634	0.7321
2	KnowTellConvince	0.7575	0.7282
3	rematchka	0.7555	0.7309
4	UL & UM6P	0.7515	0.7186
5	Itri Amigos	0.7495	0.7225
6	Raphael	0.7475	0.7221
7	Frank	0.7455	0.7173
8	Mavericks	0.7416	0.7031
9	Nexus	0.7396	0.6929
10	superMario	0.7316	0.7098
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

	Team	Micro F1	Macro F1
	Subtask 1B		
1	UL & UM6P	0.5666	0.2156
2	rematchka	0.5658	0.2497
3	AAST-NLP	0.5522	0.1425
4	Itri Amigos	0.5506	0.1839
5	HTE	0.5412	0.0979
6	Raphael	0.5347	0.1772
7	ReDASPersuasion	0.4523	0.0568
8	*Baseline (Majority)*	0.3599	0.0279
9	*Baseline (Random)*	0.0868	0.0584
10	pakapro	0.0854	0.0563

Results (Task 2)

	Team	Micro F1	Macro F1
	Subtask 2A		
1	DetectiveRedasers	0.9048	0.8626
2	AAST-NLP	0.9043	0.8634
3	UL & UM6P	0.9040	0.8645
4	rematchka	0.9040	0.8614
5	PD-AR	0.9021	0.8595
6	superMario	0.9019	0.8625
7	Mavericks	0.9010	0.8606
8	Itri Amigos	0.8984	0.8468
9	KnowTellConvince	0.8938	0.8460
10	Nexus	0.8935	0.8459
11	PTUK-HULAT	0.8675	0.7992
12	Frank	0.8163	0.6378
13	USTHB	0.7670	0.4418
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

	Team	Micro F1	Macro F1
	Subtask 2B		
1	DetectiveRedasers	0.8356	0.7541
2	UL & UM6P	0.8333	0.7388
3	AAST-NLP	0.8253	0.7283
4	rematchka	0.8219	0.7156
5	superMario	0.8208	0.7031
6	PD-AR	0.8174	0.7209
7	Itri Amigos	0.8139	0.7220
8	KnowTellConvince	0.8071	0.6888
9	USTHB	0.5046	0.1677
10	*Baseline (Majority)*	0.5046	0.1677
11	Ankit	0.4167	0.1993
12	*Baseline (Random)*	0.2603	0.2243
13	pakapro	0.2317	0.1978

Evaluation Setup

- **Development phase:** released train and development subsets, and participants submitted runs on the **development set**
- **Test phase:** participants submitted run on the official **test subset**
- **Official measure:** Micro F1

Approaches

- The most commonly used model was AraBERT, MARBERT, ARBERT, and QARiB.
- Ensembles, data augmentation, and preprocessing

Participation

Total (test phase): **20 teams** **Task 1:** 14 teams **Task 2:** 16 teams

16 teams submitted system description papers

Findings

- **Task 1 (Persuasion Technique Detection):**
  - Several participating systems showed the positive impact of exploring loss functions other than the typical Cross Entropy loss.
- **Task 2 (Disinformation Detection):**
  - We observe the systems achieved significantly high performance even in the fine-grained Subtask 2B.

Summary and Future Work

Summary

- Extended propaganda detection task with multigenre dataset (tweets + news articles)
- Disinformation detection task
- Challenges due to the skewed label distribution
- Most systems fine-tuned transformer models, used data augmentation and standard pre-processing

Future work

- Extend to multimodality of the problems
- Offer span level detection tasks

Acknowledgments

This publication was made possible by NPRP grant 14C-0916-210015. Part of this work was also funded by Qatar Foundation's IDKT Fund TDF 03-1209-210013.