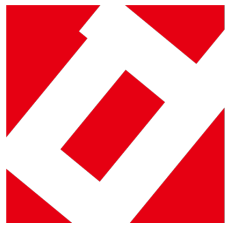
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Detecting Propaganda Techniques in Code-Switched Social Media Text

Muhammad Umar Salman, Asif Hanif, Shady Shehata, Preslav Nakov
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
{umar.salman,asif.hanif,shady.shehata,preslav.nakov}@mbzuai.ac.ae



EMNLP 2023

Background

Propaganda is a form of communication intended to influence the opinions and the mindset of the public to promote a particular agenda.

Code-switching is the mixing of multiple languages within the same text

Example: I was talking to my friend about the weather, aur phir humne decide kiya ke kal picnic pe jaenge

Propaganda Techniques
Name calling/Labeling
Repetition
Doubt
Reductio ad hitlerum
Appeal to fear/prejudice
Straw man
Loaded language
Bandwagon
Smears
Obfuscation, Int. vagueness...
Glittering generalities (Virtue)
Causal oversimplification
Appeal to authority
Red herring
Thought-terminating cliché
Black-and-white fallacy
Slogans
Whataboutism
Exaggeration/Minimisation
Flag-waving

1. Dimitrov, Dimitar, et al. "Detecting propaganda techniques in memes." *arXiv preprint arXiv:2109.08013* (2021).

Motivation

Why is there a need?

- Popularity of social media and prevalence of propaganda
- Propaganda's negative effect on people
- Presence of multilingual communities on social media
- Presence of code-switched text on social media platforms
- Need for propaganda detection systems

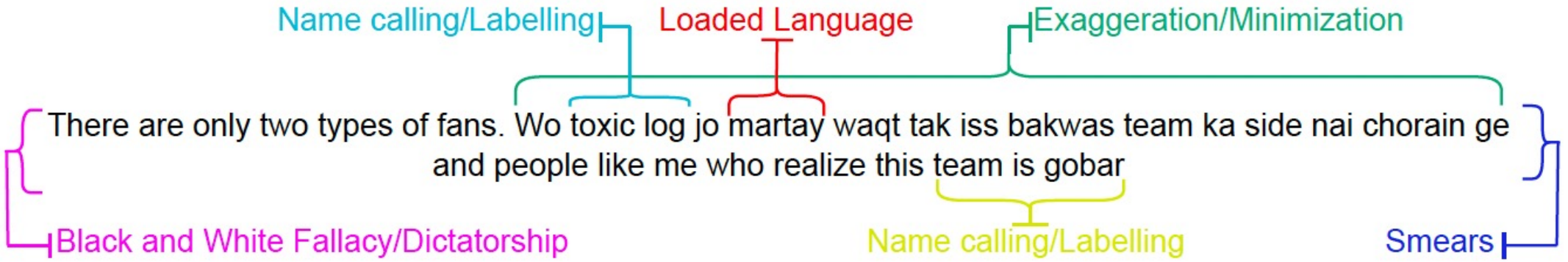
Contribution

Our Contributions

1. Formulation of novel task.

2. Creation of annotated code-switched dataset

3. Comparative analysis through running multiple experiments



Dataset

Dataset Collection

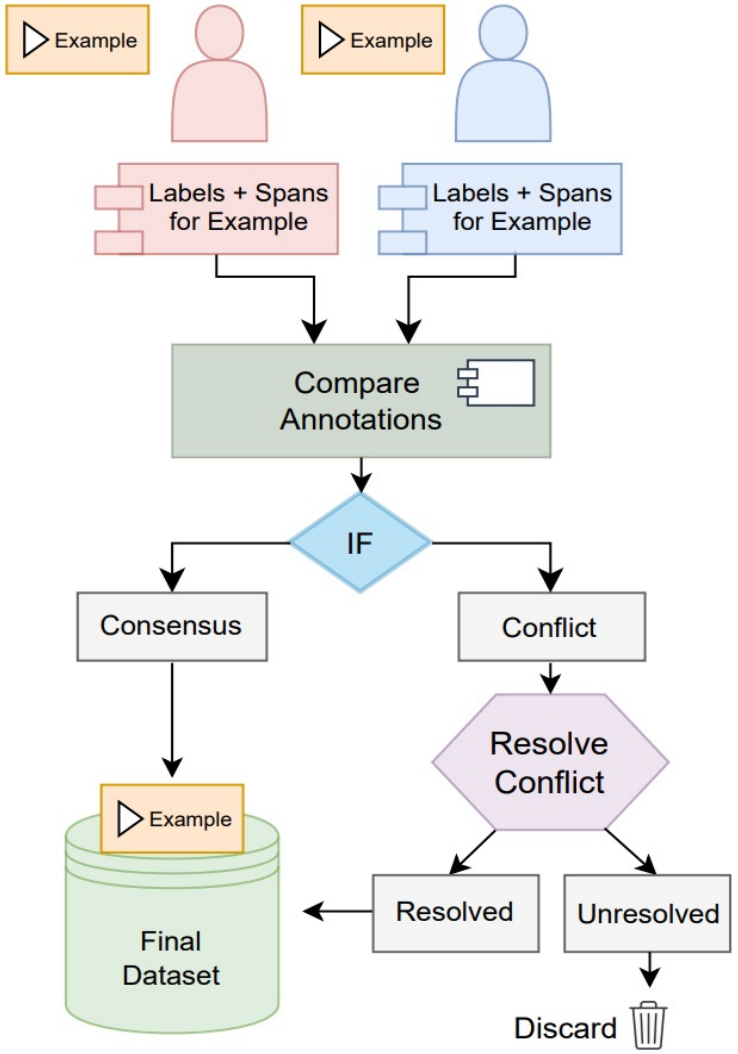
Dataset Sources: Twitter, Facebook, Instagram and YouTube

1. Four collectors (500 code-switched texts each)
2. Collectors selection criteria
3. Dataset filtering for quality purposes
4. Ethical considerations

Annotator Training

1. Details about annotators
2. Annotating propaganda techniques is a challenging task
3. Annotator’s training stages
4. Annotator’s feedback stage

Annotation Process



Quality of Annotations

Krippendorff's α [3] score between the two annotators for our annotations is 0.76

[3] Krippendorff, Klaus. "Agreement and information in the reliability of coding." *Communication methods and measures* 5.2 (2011): 93-112.

Annotation Platform

Annotate Text

Example ID #

Completed: 1030 / 1991

ID

Search

Exclude (Red)

Label Annotation Sentence

Enter Sentence Here

Prev

Next

Undo Annotation

Update Text

- Appeal to fear/prejudice
- Appeal to authority
- Whataboutism
- Slogans
- Exaggeration/Minimisation
- Loaded Language
- Smears
- Black-and-white Fallacy/Dictatorship
- Presenting Irrelevant Data (Red Herring)
- Obfuscation, Intentional vagueness, Confusion
- Name calling/Labeling
- Reductio ad hitlerum
- Bandwagon
- Repetition
- Doubt
- Flag-waving
- Causal Oversimplification
- Thought-terminating cliché
- Glittering generalities (Virtue)
- Misrepresentation of Someone's Position (Straw Man)

Label Annotation Sentence

Aurat Raj na manzoor. It is not permitted in Islam, acha lge ya bura.

Prev Next


```
{
  "id": "11",
  "text": "Aurat Raj na manzoor. It is
  "labels": [
    {
      "start_index": "0",
      "end_index": "69",
      "text_fragment": "Aurat Raj na ma
      "technique": "Black-and-white Fal
    }
  ]
}
```


Label Annotation Sentence

Aurat Raj na manzoor. It is not permitted in Islam, acha lge ya bura.

Prev Next


```
{
  "id": "11",
  "text": "Aurat Raj na manzoor. It is
  "labels": [
    {
      "start_index": "0",
      "end_index": "69",
      "text_fragment": "Aurat Raj na ma
      "technique": "Black-and-white Fal
    },
    {
      "start_index": "0",
      "end_index": "20",
      "text_fragment": "Aurat Raj na ma
      "technique": "Slogans"
    }
  ]
}
```

Label Annotation Sen

Aurat Raj na manzo
lge ya bura.

Prev

Next

lam, acha

It is

na ma

te Fal

na ma


```
{
  "id": "1",
  "text": "AMERICA WAS FOUNDED BY TOUGH HELL-RAISERS. Rugged citizens who evaded taxes,
  "labels": [
    {
      "start_index": "0",
      "end_index": "42",
      "text_fragment": "AMERICA WAS FOUNDED BY TOUGH HELL-RAISERS.",
      "technique": "Flag-waving"
    },
    {
      "start_index": "100",
      "end_index": "107",
      "text_fragment": "tyranny",
      "technique": "Loaded Language"
    },
    {
      "start_index": "152",
      "end_index": "168",
      "text_fragment": "smuggled weapons",
      "technique": "Loaded Language"
    },
    {
      "start_index": "23",
      "end_index": "41",
      "text_fragment": "TOUGH HELL-RAISERS",
      "technique": "Name calling/Labeling"
    },
    {
      "start_index": "23",
      "end_index": "41",
      "text_fragment": "TOUGH HELL-RAISERS",
      "technique": "Loaded Language"
    },
    {
      "start_index": "43",
      "end_index": "58",
      "text_fragment": "Rugged citizens",
      "technique": "Name calling/Labeling"
    }
  ]
}
```

Dataset Statistics

Statistic	Value
# of Labeled Examples	923
# of Unlabeled Examples	107
Average Example Length	147.56 ± 53.79
Maximum Example Length	400
Minimum Example Length	42
Average Span Length	63.64 ± 67.81
Maximum Span Length	400
Minimum Span Length	2
Total # of Words	28452
Vocabulary Size	7154

Table 1: Statistics about our dataset. Here the *unlabeled* examples are those with no propaganda class assigned.

Propaganda Techniques	Number of Instances	Avg Span Length ± Std Dev.
Name calling/Labeling	563	16.60 ± 10.23
Repetition	15	146.06 ± 81.02
Doubt	39	124.87 ± 76.35
Reductio ad hitlerum	8	119.00 ± 19.51
Appeal to fear/prejudice	87	144.13 ± 50.20
Straw man	18	141.38 ± 53.94
Loaded language	693	8.78 ± 8.22
Bandwagon	4	102.00 ± 19.45
Smears	382	144.25 ± 60.03
Obfuscation, Int. vagueness...	12	139.33 ± 63.04
Glittering generalities (Virtue)	44	91.97 ± 55.74
Causal oversimplification	86	90.96 ± 42.44
Appeal to authority	12	136.08 ± 63.70
Red herring	61	142.40 ± 51.66
Thought-terminating cliché	43	27.13 ± 23.03
Black-and-white fallacy	34	76.38 ± 38.95
Slogans	45	24.51 ± 12.11
Whataboutism	38	163.15 ± 56.54
Exaggeration/Minimisation	366	85.40 ± 45.23
Flag-waving	27	140.14 ± 41.14
Overall	**2,577**	**63.64 ± 67.81**

Table 2: Total number of instances and average span lengths (number of characters) for each of the propaganda classes.

Dataset Statistics

Statistic	Value
# of Labeled Examples	923
# of Unlabeled Examples	107
Average Example Length	147.56 ± 53.79
Maximum Example Length	400
Minimum Example Length	42
Average Span Length	63.64 ± 67.81
Maximum Span Length	400
Minimum Span Length	2
Total # of Words	28452
Vocabulary Size	7154

Table 1: Statistics about our dataset. Here the *unlabeled* examples are those with no propaganda class assigned.

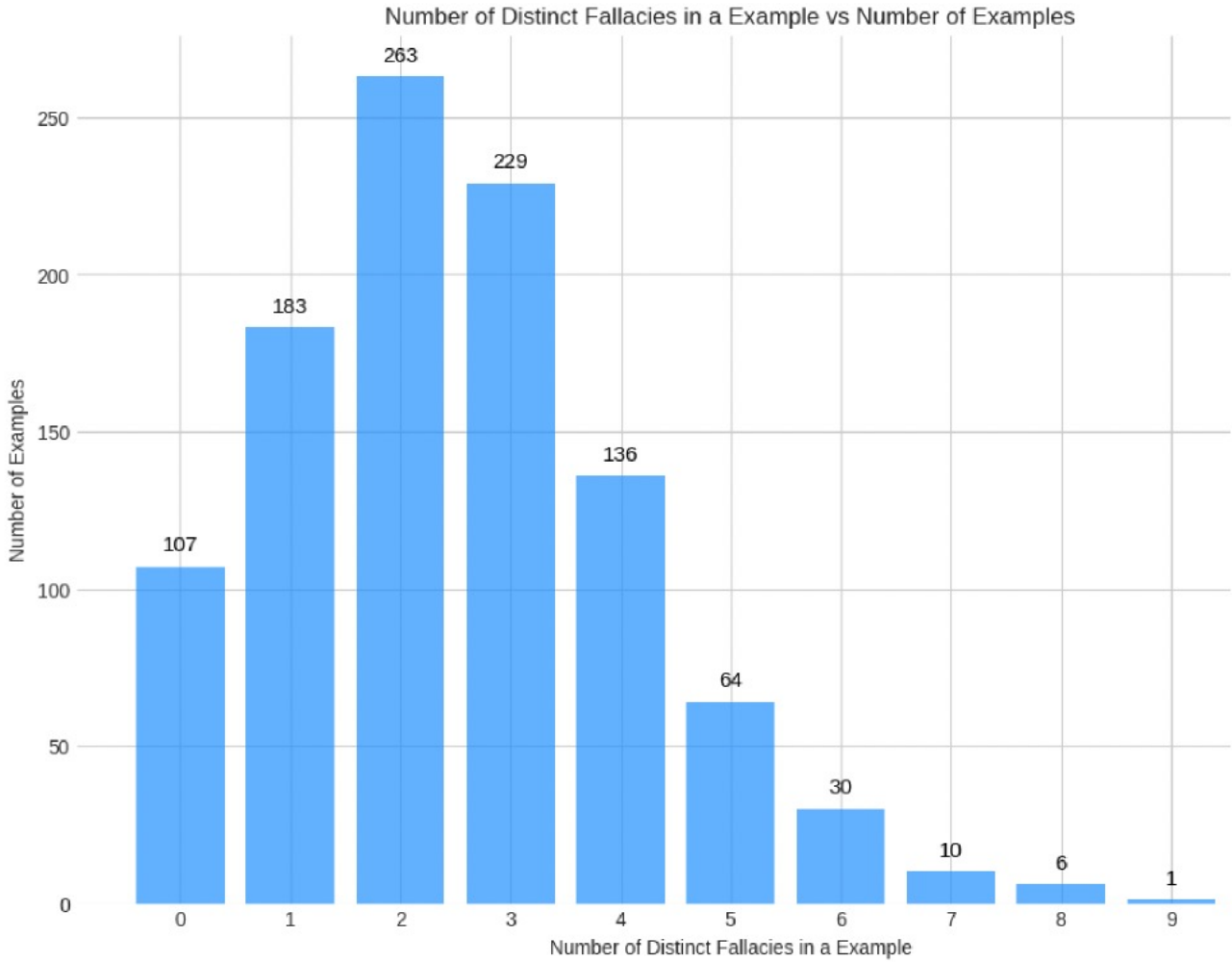


Figure 3: Histogram of the number of propaganda techniques per example.

Experiments and Results

Fine-Tuning Strategies

1. Out of Domain [1]
2. Translated (Code-switched to English) [2]
3. Code-Switched
4. No fine-tuning

Fine-Tuning Strategy Type	Model	Model Name
Out of Domain Meme Dataset (Text-Only)	$\mathcal{M}_1$	BERT
	$\mathcal{M}_2$	mBERT
	$\mathcal{M}_3$	XLM RoBERTa
Translated (Code-Switched → English)	$\mathcal{M}_4$	BERT
	$\mathcal{M}_5$	mBERT
	$\mathcal{M}_6$	XLM RoBERTa
Code-Switched	$\mathcal{M}_7$	BERT
	$\mathcal{M}_8$	mBERT
	$\mathcal{M}_9$	RUBERT
	$\mathcal{M}_{10}$	XLM RoBERTa
	$\mathcal{M}_{11}$	XLM RoBERTa (Roman Urdu)
	$\mathcal{M}_{12}$	DeBERTaV3
No fine-tuning	$\mathcal{M}_{13}$	GPT-3.5-Turbo @20-shot

Table 3: List of models and fine-tuning strategies (on particular datasets). Here $\mathcal{M}_*$ refers to a specific model fine-tuned using a specific fine-tuning *strategy type*.

1. Dimitrov, Dimitar, et al. "Detecting propaganda techniques in memes." *arXiv preprint arXiv:2109.08013* (2021).
2. https://cloud.google.com/translate/docs/

Experiment

Fine-Tuning Strategy Type	Model	Avg. Precision		Avg. Recall		Avg. F1-Score		Accuracy	Exact Match Ratio	Hamming Score
		Micro	Macro	Micro	Macro	Micro	Macro			
Out of Domain Meme Dataset (Text-Only)	$\mathcal{M}_1$	.57	.16	.18	.05	.27	.07	.898	.083	.185
	$\mathcal{M}_2$	.45	.06	.29	.07	.35	.06	.886	.071	.239
	$\mathcal{M}_3$	.44	.07	.33	.08	.39	.07	.889	.083	.261
Translated (Code-Switched → English)	$\mathcal{M}_4$	.45	.12	.44	.12	.44	.10	.884	.038	.288
	$\mathcal{M}_5$	.49	.10	.37	.11	.42	.10	.891	.064	.267
	$\mathcal{M}_6$	.54	.26	.40	.14	.46	.16	.900	.103	.320
Code-Switched	$\mathcal{M}_7$	.55	.21	.37	.12	.44	.14	.900	.096	.308
	$\mathcal{M}_8$	.50	.24	.32	.12	.39	.14	.893	.083	.263
	$\mathcal{M}_9$	.49	.10	.35	.09	.40	.10	.892	.083	.280
	$\mathcal{M}_{10}$	.54	.21	.43	.16	.48	.17	.901	.110	.354
	$\mathcal{M}_{11}$	**.59**	.34	.49	.22	**.53**	.25	**.910**	**.135**	**.375**
	$\mathcal{M}_{12}$	.51	**.53**	.43	.15	.46	.17	.895	.090	.307
No fine-tuning	$\mathcal{M}_{13}$	.39	.31	**.53**	**.42**	.45	**.28**	.862	.051	.306

Table 4: Results on the nine evaluation measures listed in subsection 4.4 for the different models $\mathcal{M}_1$ to $\mathcal{M}_{13}$. Green highlights show the highest score for each of the evaluation measures.

Experiment

Models → Propaganda Techniques ↓	Percentage of Instances (%)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$	$\mathcal{M}_9$	$\mathcal{M}_{10}$	$\mathcal{M}_{11}$	$\mathcal{M}_{12}$	$\mathcal{M}_{13}$
Loaded Language	26.9	.52	.61	.63	.60	.57	.66	.64	.63	.61	**.74**	.70	.70	.63
Obfuscation, Intentional vagueness, Confusion	0.50	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Appeal to fear/prejudice	3.40	.00	.00	.00	.12	.00	.30	.20	.30	.00	**.35**	.30	.32	.33
Appeal to authority	0.50	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.15
Whataboutism	1.50	.00	.00	.00	.00	.14	**.40**	.00	.25	.00	.00	.20	.00	.40
Slogans	1.70	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.18**
Exaggeration/Minimisation	14.2	.00	.00	.00	.25	.17	.29	.40	.31	.44	.47	**.56**	.34	.37
Black-and-white Fallacy/Dictatorship	1.30	.00	.00	.00	.00	.00	.00	.29	.25	.10	.00	.33	.33	**.55**
Smears	14.8	.22	.09	.27	**.59**	.57	.57	.47	.47	.48	.49	.53	.48	.40
Doubt	1.50	.29	.00	.00	.00	.00	.00	.29	.00	.00	.40	**.50**	.22	.31
Bandwagon	0.20	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.50**
Name calling/Labeling	21.8	.32	.46	.52	.51	.57	.52	.52	.32	.45	.51	.63	.56	**.69**
Reductio ad hitlerum	0.30	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.12**
Presenting Irrelevant Data (Red Herring)	2.40	.00	.00	.00	.00	.00	**.20**	.00	.20	.00	.00	.00	.00	.00
Repetition	0.60	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.33**
Straw Man	0.70	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Thought-terminating cliché	1.70	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.22**	.00	.00
Glittering generalities (Virtue)	1.70	.00	.00	.00	.00	.00	**.29**	.00	.00	.00	.25	.20	.22	.20
Flag-waving	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.36**	.00	.22
Causal Oversimplification	3.30	.00	.00	.00	.00	.00	.00	.00	.13	.00	.22	**.50**	.12	.24

Table 5: Comparison of class-level performance (F1-Score) on 13 different models. The naming convention of the models can be found in Table 3. Green highlights indicate the highest F1-Score for each propaganda technique.

Conclusion & Future Work

Conclusion

1. Novel task detecting propaganda techniques in code-switched text

2. Created corpus of 1030 code-switched annotated with 20 propaganda techniques

3. Run preliminary experiments using different fine-tuning strategies and models

4. Find modelling multilinguality rather than using translation is more effective for our task

Future Work

1. Expand our annotated corpus with many more examples

2. Run our experiments for our task for other resource-poor languages and compare results

3. See how different models and fine-tuning strategies perform on detecting propaganda on a fragment level

4. We want to understand the best strategy for handling codeswitched text with LLMs

Thank you