EMNLP 2023

# 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

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Github Repo

## Key Contribution

We propose a novel task of detecting propaganda techniques in code-switched texts. For this, we create an annotated corpus at a fragment level. Our analysis confirms that directly modeling multilinguality outperforms the translation of code-switched text for the task of propaganda detection.

Name calling/Labelling | Loaded Language | Exaggeration/Minimization

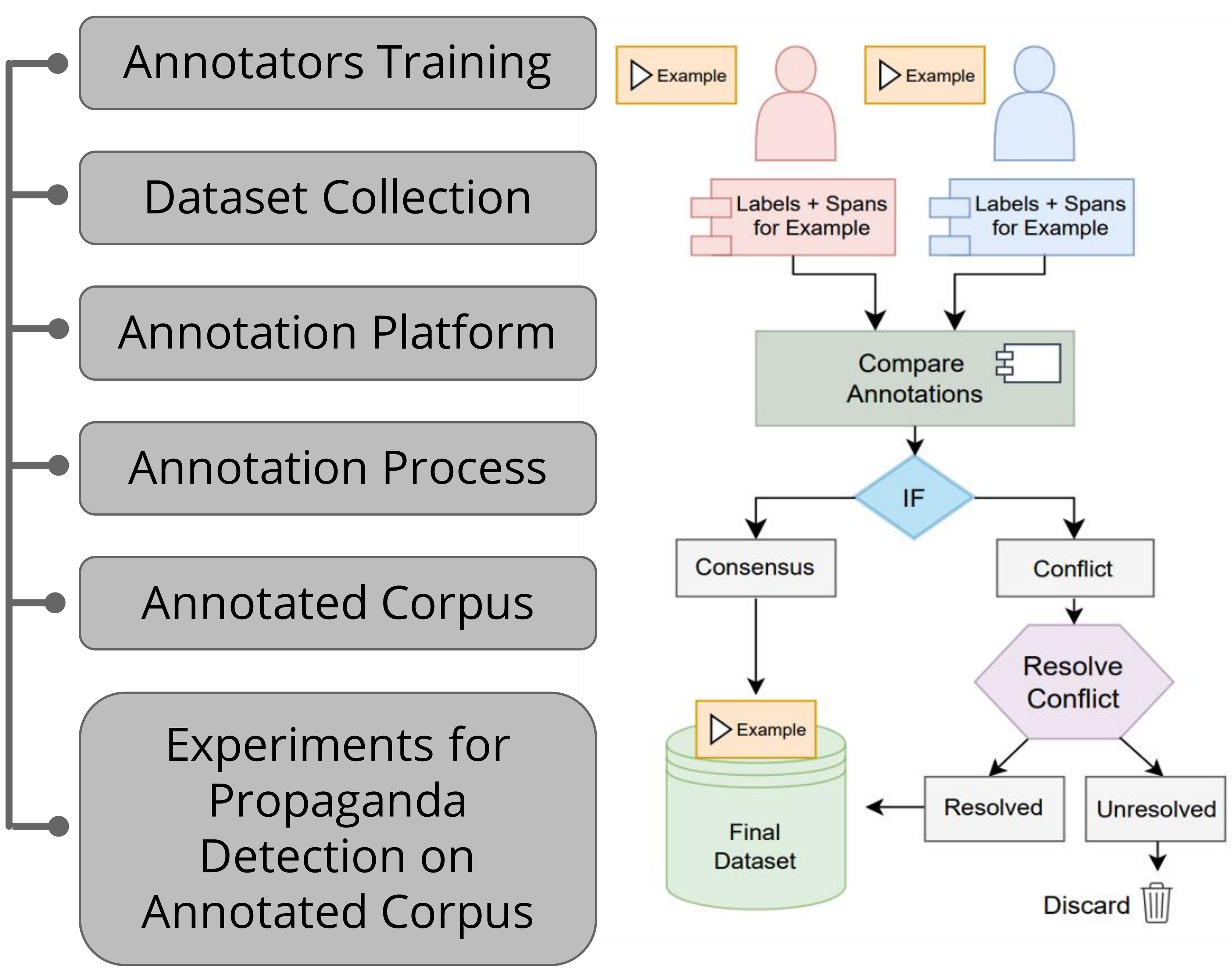
There are only two types of fans. Wo toxic log jo martay waqt tak iss bakwas team ka side nai chorain ge and people like me who realize this team is gobar

Black and White Fallacy/Dictatorship | Name calling/Labelling | Smears

## Introduction

- Propaganda is the dissemination of misleading information in order to manipulate a target audience's opinion towards a particular objective
- The influence of social media platforms, coupled with accessing news information through them, has led to a swift and widespread dissemination of propaganda
- Existing propaganda detection efforts primarily focus on high-resource languages, neglecting low-resource languages as well as code-switched text.
- Social media platforms are used by millions of multilingual users which resort to mixing multiple languages (code-switching) on such platforms.

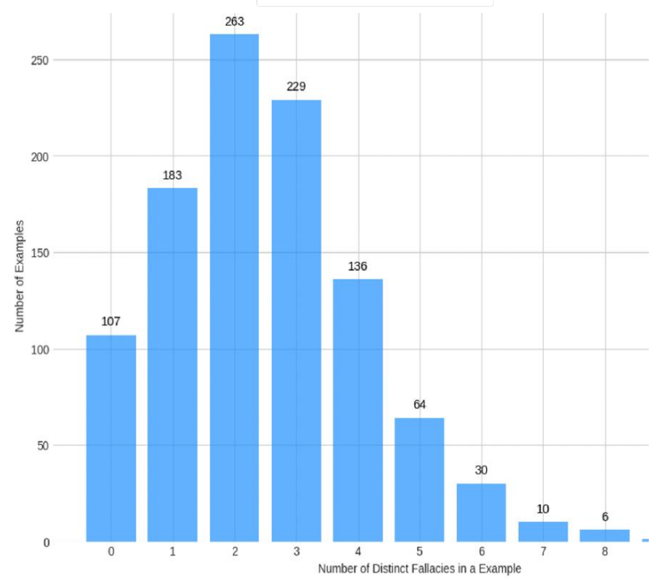
## Dataset and Annotation





| 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 |

| Agreement Pair | Krippendorff's $\alpha$ |
|---|---|
| Annotator 1 vs. Annotator 2 | 0.76 |




| 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** |

## Experiments and Results

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

| Fine-Tuning Strategy Type | Model | Avg. Precision Micro | Avg. Precision Macro | Avg. Recall Micro | Avg. Recall Macro | Avg. F1-Score Micro | Avg. F1-Score Macro | Accuracy | Exact Match Ratio | Hamming Score |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 1: Results on the 9 evaluation measures for the different models $\mathcal{M}_1$ to $\mathcal{M}_{13}$. Green highlights show the highest score for each of the evaluation measures.

| 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 2: Comparison of class-level performance (F1-Score) on 13 different models. Green highlights indicate the highest F1-Score for each propaganda technique.

## Conclusion & Future Work

1. Novel task of propaganda detection
2. Created corpus of 1030 code-switched texts
3. Comparative analysis using fine-tuning strategies and models.
4. Find modelling multilinguality rather than using translation is more effective for our task

1. Expansion of annotated corpus
2. Experiments for other resource-poor languages
3. Experiments to detect propaganda at a fragment level
4. Exploring better alternatives to handle propaganda detection on codeswitched text with LLMs