



The 4rd Italian Conference on Big Data and Data Science
University of Cagliari

**Exploring the Dataset Landscape for
Automated Propaganda Detection: A
Data-Centric Insight**

Introduction

- ❑ Propaganda is an intentional and systematic form of communication
- ❑ The main aims is to influence opinions and beliefs of an audience
 - ❑ selection, omission or distortion of information
 - ❑ employment of persuasive techniques (emotional appeals, oversimplification, slogans)
- ❑ The ability to discern propaganda is significant
- ❑ Reduce the risk of manipulation
- ❑ Promote critical thinking

Our work

- ❑ Mini review - 10 research articles - 2021/2025
- ❑ Focus on data rather than methods
- ❑ Select only span level annotations
 - ❑ more fine-grained and interpretable analysis
- ❑ We analyse dataset, annotation schema, methodology

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum **omnium fortissimi** sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe comeant atque ea **quae ad effeminandos animos pertinent important**, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa **Helvetii quoque reliquos Gallos virtute praecedunt**, quod fere **cotidianis proeliis cum Germanis contendunt**, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language

Attack on Reputation: Smears

Manipulative Wording: Exaggeration

Justification: Appeal to Values



Commissio
PopulusQue
Europaea

Dataset Overview 1

- ❏ PTC - 2019 - 433 cits.
- ❏ SemEval2020 - 2020 - 255 cits.
- ❏ SemEval2023 - 2023 - 123 cits.
- ❏ WANLP - 2022 - 77 cits.

Dataset Overview - PTC

- ❑ Propaganda Techniques Corpus
 - ❑ Dataset for fine-grained propaganda detection
- ❑ Contains 451 English news articles
 - ❑ Span level annotations
- ❑ Each fragment is labelled with one of 18 propaganda techniques
- ❑ Annotations contains
 - ❑ Precise character offsets
 - ❑ Spans annotated by experts

Dataset Overview - PTC - SemEval 2020

- ❑ Proposed for the SemEval-2020 Task 11
- ❑ Focuses on detecting propaganda techniques in English articles
- ❑ Two subtask:
 - ❑ span identification
 - ❑ techniques classification
- ❑ Contains over 500 news articles

Dataset Overview - PTC - SemEval 2023

- ❑ Similar to SemEval 2020
- ❑ Main novelty: six languages
 - ❑ English, French, German, Italian, Polish and Russian
- ❑ Resources collected from 2020 to 2022
- ❑ Wide range topics:
 - ❑ from COVID-19 to Russo-Ukrainian war
- ❑ Data draws from mainstream and alternative media
 - ❑ many flagged as potential disinformation spreaders

Dataset Overview - WANLP

- ❑ Developed for WANLP 2022 Shared Task 3
 - ❑ Propaganda Detection in Arabic Social Media Text
- ❑ Consists of Arabic news articles
- ❑ Propaganda techniques span level annotated
- ❑ Inspired by PTC
 - ❑ First dataset of this kind in Arabic

Dataset Overview - 2

- ❑ BanMANI, 800 Bangla social-media posts
- ❑ Salman et al., 1030 English-Roman Urdu social-media posts
- ❑ ArAIEval, 9000 Arabic text snippet (between tweets and news)
- ❑ ArMPro, 20k annotated Arabic spans; 23 propaganda tech.
- ❑ ZenPropaganda, 125 texts in Russian focus on COVID-19; 36 propaganda tech.
- ❑ ManiTweet, 3600 tweet-article pairs

Key Factors

- ❑ Class imbalance
 - ❑ Loaded Language, Name Calling are overrepresented
 - ❑ Straw Man, Whataboutism underrepresented (near-zero F1 scores)
- ❑ Annotation Schemas
 - ❑ Degree of granularity and definition vary considerably
 - ❑ Many datasets follow SemEval 2020 Task 11
 - ❑ ZenPropaganda defines 36 distinct techniques
 - ❑ ArMPro set 23 techniques and added “no tech” for neutral span

Key Factors

❑ Linguistic diversity

- ❑ Many datasets are English-centric
- ❑ Expansion to Russian, Arabic, Polish and Bangla
- ❑ LLMs limitations in zero-shot

❑ Domain coverage

- ❑ Mainly news media consisting of formal text like SemEval 2020/2023 and ArMPro
- ❑ ArMPro focus on Arabic news
- ❑ ZenPropaganda center the scope on Russian COVID news
- ❑ Social media present more challenging environment
- ❑ Short, informal and noisy content

Key Factors

- ❑ No DCAI paradigm
 - ❑ None datasets provide life-cycle maintenance or regular update
 - ❑ Propaganda techniques, languages and media evolve
 - ❑ Models trained on outdated data
 - ❑ risk of poor generalization
 - ❑ reduce robustness
 - ❑ weak to new manipulation strategies
- ❑ SemEval and PTC seems moving in that direction
- ❑ Nevertheless, no regular update or periodically re-annotation or augmentation are announced

Methodologies

- ❑ Deep Learning Models
 - ❑ BERT, RoBERTa, XLM-RoBERTa
 - ❑ Data augmentation
- ❑ Large Language Models
 - ❑ GPT-3.5, GPT-4
 - ❑ Chain of thought, few-shot learning
- ❑ Limitations:
 - ❑ class imbalance
 - ❑ irreproducibility
 - ❑ hallucination
 - ❑ topic-specific

Conclusions

- ❑ Pervasive issue of class imbalance
- ❑ Lack of standardization in annotations
- ❑ Need for regular update and continuous improvement

- ❑ Future directions:
 - ❑ develop richer, multilingual, cross-domain datasets
 - ❑ incorporations of multimodal data (text and images)

Thanks for your
attention