Oppositional Thinking Analysis: Conspiracy vs Critical Narratives
PAN CLEF 2024 Shared Task

Damir Korenčić, Berta Chulvi, Xavier Bonet
Mariona Taule, Paolo Rosso, Francisco Rangel

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UNIVERSITAT DE BARCELONA
symanto psychology ai

presenter: Damir Korenčić, PhD
PRHLT Research Center, Universitat Politècnica de València

Outline

1. Background & Motivation
2. Corpus & Annotation
3. Subtask1 - Distinguishing Between Critical and Conspiracy
4. Subtask2 - Detecting Elements of the Oppositional Narratives
5. Challenge Guidelines

Conspiracy Texts vs Critical Texts

Conspiracy Theories (CTs) are complex narratives that attempt to explain the ultimate causes of significant events as cover plots orchestrated by secret, powerful, and malicious groups (Douglas et al. 2023)

Critical texts oppose mainstream views of events. For example, related to the pandemic – efficacy/safety of vaccines and/or public health policies.

Conspiracy Texts vs Critical Texts

Existing approaches do not distinguish between critical and conspiratorial thinking. This distinction is important because labeling a text as conspiratorial when it is, in fact, oppositional to mainstream views, could potentially lead those who were simply asking questions closer to conspiracy communities (Sutton et al. 2022; Funkhouser 2022).

If the models do not differentiate between critical and conspiratorial thinking, there is a ***high risk of pushing people toward conspiracy communities****.*

Conspiracy Texts vs Critical Texts

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

- Conspiracy texts: supporting/suggesting/implying conspiracy theories
- Critical texts: oppose mainstream views of events, but are not conspiracies
- **Oppositional texts**: critical or conspiracy text

EMERGENCY BROADCAST: Globalists Set Stage for Cashless Dystopia with Full - Spectrum Surveillance as Deep State Scrambles to Bury Truth on Deadly Covid Jabs.

Another Victim of the nano meta antenna. Injected & Chipped! Graphene ferrous oxide. Nano lipid particles. 5G SMART - Secret Militarized - Armaments - Residential Technologies SMART The human race is being turned into the Internet of bodies.

Cardiologist says likely contributory factor to excess cardiovascular deaths is the COVID mRNA vaccine and roll out should be suspended pending an inquiry.

I'm deeply concerned that the push to vaccinate these children is nothing more than a dystopian experiment with unknown consequences.

Elements of Oppositional Narratives

- Agents – responsible for the actions and/or negative effects
- Facilitators – those who help the agents
- Victims – suffer the consequences of the actions of the agents
- Campaigners – those who oppose the mainstream narrative
- Objectives – the intentions of the agents
- Effects – the negative consequences suffered by the victims

Computational analysis of conspiratorial texts fails to address is the role that intergroup conflict (IGC) (Böhm et al. 2020) plays in these narratives. ***Intergroup conflict*** *is a way of framing events by emphasizing the hostility between groups, typically by using "us versus them" narrative, and by fueling the perceived injustice and threat to the group.*

The increasing potentially violent involvement of conspiracist communities in political processes suggests that one of the purposes of CTs is to enforce IGC and coordinate action (Wagner-Egger et al. 2022). Therefore, tools that enable an IGC-based analysis of conspiratorial texts could offer valuable insights for content moderation.

The proposed scheme identifies the following categories: "facilitators" (collaborators of the agents, such as the media) and "campaigners" (those that unmask the conspiracy agenda).

These are "key players" in IGC: the facilitators are tangible targets with whom real conflict is possible, and the campaigners are those that show their opposition to the facilitators and try to persuade the victims to join their cause.

Elements of Oppositional Narratives

- Agents – responsible for the actions and/or negative effects
- **Facilitators** – those who help the agents
- Victims – suffer the consequences of the actions of the agents
- **Campaigners** – those who oppose the mainstream narrative
- Objectives – the intentions of the agents
- Effects – the negative consequences suffered by the victims

Elements of Oppositional Narratives

Private owned WHO **A** with investors like Bill Gates **A** can declare a new pandemic out of thin air anytime they want and the world governments ruled by their puppets **F** as well as their media **F** starts with the constant fear mongering **E** , getting people **V** to get their pharma companies **A** injections and drugs that are magically ready in light speed, clear induction that they have been ready for the orchestrated fake pandemics, long before they start with the constant fear mongering **E** by the media **F** and governments **F** . To those awake already **C** , we know their games and agenda **O** , but sadly most people **V** fall for it, again and again and pay a hefty price, often with their health, lives, the loss of their loved ones **E** . These are very evil beings **A** , intent on destroying us **O** regular people **V** .

Agents (A), Facilitators (F), Campaigners (C), Victims (V)
Effects (E), Objectives (O)

The Corpus

- Telegram texts related to COVID-19
- List of oppositional Telegram channels
- Keyword-filtered for COVID-19
- Quality index (channel activity, num. of keywords, . . .)
- 5.000 English, 5.000 Spanish texts

- 3 annotators per text, averaging
- High IAA: 0.86 F1 (English), 0.89 F1 (Spanish)

Language	Conspiracy	Critical
English	1724 (34.48%)	3276 (65.52%)
Spanish	1828 (36.56%)	3172 (63.44%)

Table: Per-language class distribution

- 2 annotators per text, merging the labels
- Good IAA: 0.718 gamma

		A	F	C	V
es	All	3329 (14.0%)	2688 (11.3%)	4231 (17.8%)	5260 (22.2%)
	Con.	1361 (9.8%)	1184 (8.6%)	2133 (15.4%)	3543 (25.6%)
	Crit.	1968 (20.0%)	1504 (15.2%)	2098 (21.3%)	1717 (17.4%)
en	All	6411 (22.4%)	3462 (12.1%)	6416 (22.4%)	4433 (15.5%)
	Con.	3333 (21.1%)	1336 (8.5%)	3839 (24.4%)	2734 (17.3%)
	Crit.	3078 (23.9%)	2126 (16.5%)	2577 (20.0%)	1699 (13.2%)

		O	E
es	All	622 (2.6%)	7150 (30.2%)
	Con.	23 (0.2%)	5326 (38.5%)
	Crit.	599 (6.1%)	1824 (18.5%)
en	All	2073 (7.2%)	5565 (19.4%)
	Con.	615 (3.9%)	3708 (23.5%)
	Crit.	1458 (11.3%)	1857 (14.4%)

Table: Number of spans across languages and categories

- Binary classification: 'Critical' vs. 'Conspiracy'
- Evaluation: MCC (Chicco et al. 2020), per-class F1 scores, macro-averaged F1

- BERT (Devlin et al. 2019; Cañete et al. 2023)
- Hyperparameters:
  LR 2×10^{-5}, batch size 16, num. epochs 3, warmup 10%
- Evaluation: 5-fold crossvalidation

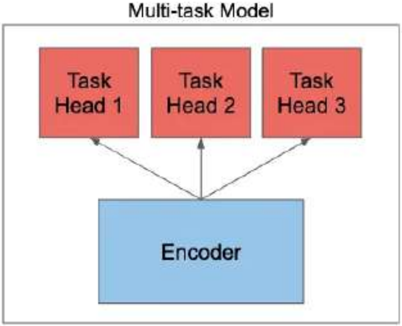
Table: Baseline classifiers' performance

Language	MCC	F1-Consp	F1-Crit	F1-avg
English	0.754	0.837	0.916	0.876
Spanish	0.690	0.801	0.888	0.844

Sequence Labeling Problem

- Six span categories:
  Agents, Facilitators, Victims, Campaigners, Objectives, Effects
- Long and **overlapping spans**
- Evaluation: span-F1 (Da San Martino, Yu, et al. 2019)

- Multi-task learning (Ruder 2017)
- BERT (Devlin et al. 2019; Cañete et al. 2023)
- Token-classification layer (BIO scheme)
- Hyperparameters:
  LR 2×10^{-5}, batch size 16, num. epochs 10, warmup 10%
- Evaluation: 5-fold crossvalidation, span-F1



Baseline Approach

Table: Baseline span-F1 scores by language and category

Lang.	Cum.	A	F	V	C	O	E
English	0.488	0.610	0.384	0.598	0.560	0.444	0.576
Spanish	0.478	0.475	0.383	0.571	0.499	0.312	0.636

Scores on a (relatively) similar propaganda detection task:
0.52 span-F1 (Da San Martino, Barrón-Cedeño, et al. 2020)

The Challenge

High-level Guidelines

- Avoid making only trivial tweaks to the baselines
- Try to answer a research question:
  what is the performance an approach X on this task?
  why does it work/not work?
- Try to analyze model performance
- Robust performance assessment
- Working notes papers with good results
  but a weak "conceptual background" will not be smiled upon
- Data analysis is a bonus

Hyperparameter Optimization

- Baselines: standard hyperparameters
- Grid search
- Bayesian optimization (Wu et al. 2019)
- Regular vs. nested cross-validation (Cawley et al. 2010)
- Validation set (less training data)
- Computationally intensive

- Freezing the layers
- Adaptive learning rate (Howard et al. 2018)

- Ensembles
- Choice of the transformer
- Non-transformer methods (Kowsari et al. 2019)

- Choice of the transformer
- Improving the multi-task learning (Ruder 2017; Worsham et al. 2020)
  - Task definition
  - Algoritmhic methods: loss balancing, ...

- Curse of multilinguality (Pfeiffer et al. 2022)
- Choice of the transformer
- Translation?

Sequence-to-sequence models

- T5 (Raffel et al. 2020)
- Multi-task learning
- Sequence-labeling 'coding'
- Computational costs

- TweetBERT, CT-BERT (Qudar et al. 2020; Müller et al. 2023)
- Model adapted for Telegram text?

LLMs

- Few-shot or zero-shot learning
- Prompt instability (Zhao et al. 2021)
- Sequence labeling (Wang et al. 2023)
- Data augmentation
  - Classification: text rephrasing
  - Sequence labeling: need to tackle the annotations
- Feature extraction

GitHub Repository

- https://github.com/dkorenci/pan-clef-2024-oppositional
- Use the baseline code
- Data in spaCy format

Bibliography I

Douglas, Karen M. et al. (2023). “What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication”. In: *Annual Review of Psychology* 74.1, pp. 271–298.

Sutton, Robbie M. et al. (2022). “Rabbit Hole Syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs”. In: *Current Opinion in Psychology* 48, p. 101462.

Funkhouser, Eric (2022). “A tribal mind: Beliefs that signal group identity or commitment”. In: *Mind & Language* 37.3, pp. 444–464. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mila.12326.

Bibliography II

Böhm, Robert et al. (2020). "The psychology of intergroup conflict: A review of theories and measures". In: *Journal of Economic Behavior & Organization* 178, pp. 947–962.

Wagner-Egger, Pascal et al. (2022). "Awake together: Sociopsychological processes of engagement in conspiracist communities". In: *Current Opinion in Psychology* 47, p. 101417.

Chicco, Davide et al. (Jan. 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC Genomics* 21.1, p. 6.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Cañete, José et al. (2023). *Spanish Pre-trained BERT Model and Evaluation Data*. ArXiv:2308.02976. arXiv: 2308.02976 [cs.CL].

Da San Martino, Giovanni, Seunghak Yu, et al. (Nov. 2019). "Fine-Grained Analysis of Propaganda in News Articles". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5636–5646.

Ruder, Sebastian (June 2017). *An Overview of Multi-Task Learning in Deep Neural Networks*. arXiv:1706.05098.

Bibliography V

- Da San Martino, Giovanni, Alberto Barrón-Cedeño, et al. (2020). "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 1377–1414.
- Wu, Jia et al. (2019). "Hyperparameter optimization for machine learning models based on Bayesian optimization". In: *Journal of Electronic Science and Technology* 17.1, pp. 26–40.
- Cawley, Gavin C et al. (2010). "On over-fitting in model selection and subsequent selection bias in performance evaluation". In: *The Journal of Machine Learning Research* 11, pp. 2079–2107.

Bibliography VI

Howard, Jeremy et al. (2018). "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339.

Kowsari, Kamran et al. (Apr. 2019). "Text Classification Algorithms: A Survey". In: *Information* 10.4. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 150.

Worsham, Joseph et al. (Aug. 1, 2020). "Multi-task learning for natural language processing in the 2020s: Where are we going?" In: *Pattern Recognition Letters* 136, pp. 120–126.

Pfeiffer, Jonas et al. (2022). "Lifting the Curse of Multilinguality by Pre-training Modular Transformers". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3479–3495.

Raffel, Colin et al. (July 28, 2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv: `1910.10683[cs,stat]`.

Qudar, Mohiuddin Md Abdul et al. (2020). "Tweetbert: a pretrained language representation model for twitter text analysis". In: *arXiv preprint arXiv:2010.11091*.

Müller, Martin et al. (2023). "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter". In: *Frontiers in Artificial Intelligence* 6.

Zhao, Zihao et al. (July 1, 2021). “Calibrate Before Use: Improving Few-shot Performance of Language Models”. In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, pp. 12697–12706.

Wang, Shuhe et al. (May 12, 2023). *GPT-NER: Named Entity Recognition via Large Language Models*. arXiv: 2304.10428[cs].