

XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (*)

Paolo Rosso



Mariona Taulé



David Camacho



David Arroyo



Juan Gómez




UNIVERSIDAD
DE GRANADA

Francisco Rangel
symanto
psychology ai

SEPLN-CEDI A Coruña, 20/06/24

Outline

- The problem of disinformation in Spain and in Europe
- Disinformation in XAI-DisInfodemics
- Conspiracy theories in XAI-DisInfodemics
- The shared task at  **PAN** on Oppositional thinking analysis:
Conspiracy theories vs critical thinking narratives

Disinformation in Spain

- 88% of Spanish citizens consider that disinformation is a problem

Eurobarometer 464, April 2018: **Fake news** and **disinformation** online

https://data.europa.eu/euodp/es/data/dataset/S2183_464_ENG

- 66% of them come across to **false information** at least once a week

Eurobarometer 503, March 2020: Attitudes towards the impact of digitalisation on daily lives

https://data.europa.eu/euodp/es/data/dataset/S2228_92_4_503_ENG

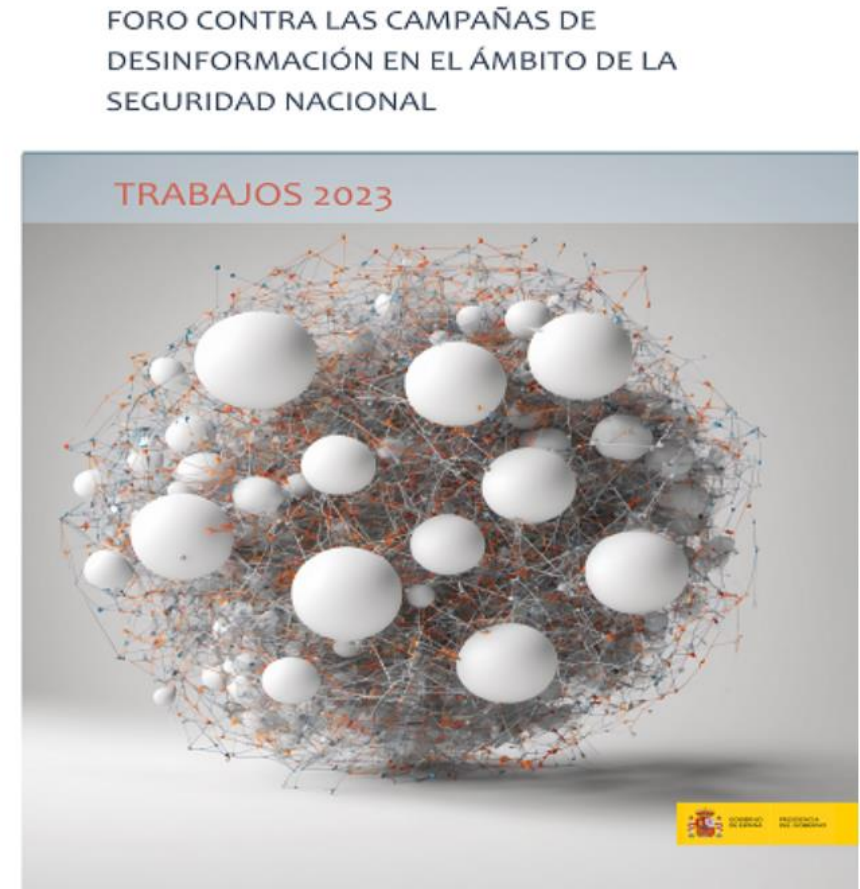
- 86% of Spanish citizens consider that **disinformation** changes the reality and is a problem for the country; 78% comes across often to **false information** vs 69% on average in EU

Eurobarometer 98, Winter 2022/2023

<https://europa.eu/eurobarometer/surveys/detail/2872>

National Department of Security

- Book by the **Spanish National Department of Security** on Disinformation campaigns
- **Information warfare:** foreign information manipulation interference (e.g. from Russia)
- **Chapter on AI to fight disinformation:**
 - Machine learning and deep learning AI techniques
 - Natural language processing
 - Social network analysis
 - The role of fact-checkers
 - Deepfakes
 - Large Language Models (LLMs) and the automatic generation of texts



<https://www.dsn.gob.es/es/documento/foro-contra-campa%C3%B1as-desinformaci%C3%B3n-%C3%A1mbito-seguridad-nacional-trabajos-2023>

Congress of Deputies

- Report by **Oficina-C**,
the Office of Science and Technology of the
Spanish Congress of Deputies

oficinac.es/informes-c/desinformacion-era-digital

Report C

Disinformation in the digital age

A complex threat for democracies

Summary C	1	Impact	16
Introduction	3	Combatting disinformation: agents and mitigating strategies	17
An evolving conceptual framework	3	Guarantees, detection and neutralization	18
Disinformation and other information disorders	3	Automation: artificial intelligence as an ally	19
Narratives for disinformation	5	Progress in regulations	22
Scope and relevance in the digital age	5	Privacy, security and elections	25
A new social and informational context	6	A strategic, participative vision of the future	25
Agent: instigators and distributors	7	Key concepts	26
Channels: digital impact and prevalence of classic channels	7	Bibliography:	I
Content	8		
Receiver	8		
Contemporary phenomena involved in the rise of disinformation	9		
Trust and the democratic framework	9		
Information mediation and journalism	9		
Social fragmentation	10		
Cognition and individual vulnerability	10		
Digital governance and business models	12		
Technologies that can be used for disinformation			

EDMO Task Force on 2024 EU Elections

EDMO Task Force On 2024 European Elections



Cloacked science



Dr. Li-Meng YAN ✓

@DrLiMengYAN1

MD PhD
Independent virologist

Cloacked Science uses scientific jargon to hide agendas under legitimate appearances. **Dr. Joan Donovan** coined the term to understand the disinformation campaign claiming SARS-CoV-2 was lab-engineered. Despite being debunked, these reports lent credibility to conspiracy theories about COVID-19.

Assessment of the veracity of the claim

1. Create a data lake of credible sources (Media Bias Fact Check, AllSides or Ad Fontes Media)
2. Name entity recognition and linking (**NER** + **NEL**) to extract cited experts and organizations.
3. Credibility Check:
 - Determines if cited sources are reputable.
 - Flags researchers cited out of their expertise.

CO2 is making Earth greener—for now

By Samson Reiny,
NASA's Earth Science News Team



Boston University, ORG; NASA, ORG; Nature Climate Change, ORG; Peking University, ORG

CO2 Has Almost No Effect on Global Temperature, Says Leading Climate Scientist

BY CHRIS MORRISON 24 SEPTEMBER 2022 5:48 PM

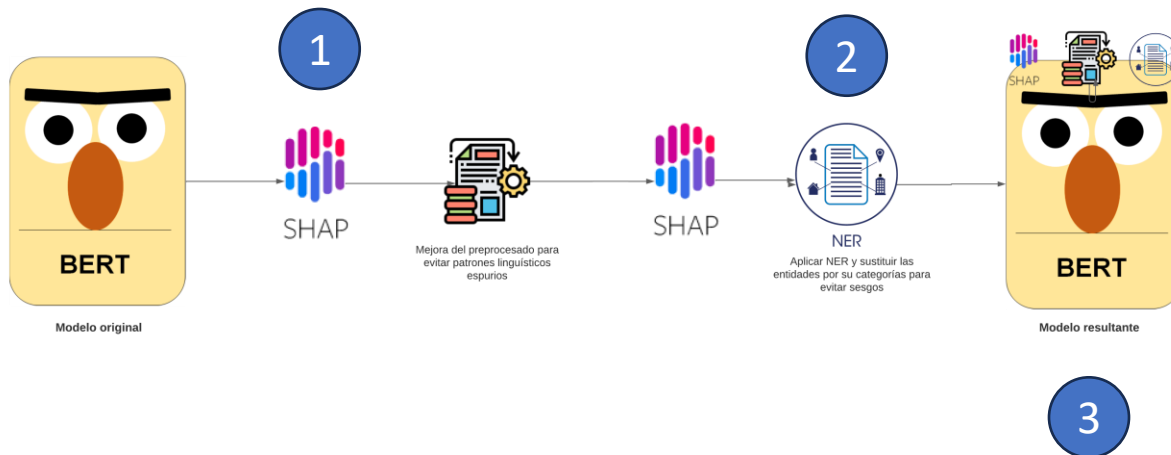


Australian Government, ORG; BBC, ORG; BP, ORG; Chris Morrison, PER; Global Warming Policy Foundation, ORG; William Kininmonth, PER; ...

Arroyo D., Degli Esposti S., Gómez A., Palmero S., Pérez L. (2023) **On the Design of a Misinformation Widget (MsW) against Cloacked Science**, In: S. Li, M. Manulis, A. Miyaji (Eds.), Network and System Security, Lecture Notes in Computer Science, Springer Nature Switzerland, pp. 385–396

Enhancing disinformation detection with eXplainable AI and Named Entity replacement

XAI methods to improve generalization in classification by anonymizing named entities
[UGR + Eugenio Martínez et al. @ UJA]



- (1) **SHAP** method (**SHapley Additive exPlanations**) is used to identify segments more relevant to the model
- (2) Replace named entities with placeholders in preprocessing (disinformation is frequently targeted to people, organizations and locations)
- (3) Training and validation metrics decrease for the train dataset but increase for an external dataset

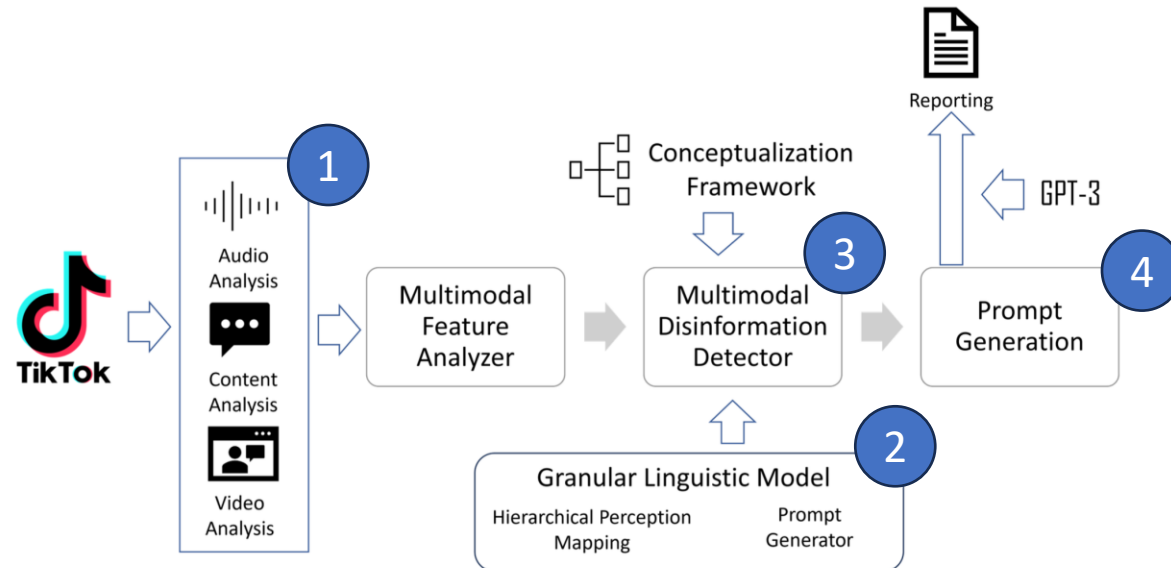
Bolivia approved the use of chlorine dioxide amid the fight against covid-19.

Named Entity
Overrepresentation

This classifier will tag any new sentence with "Bolivia" as **false**!

An intelligent approach for multimodal disinformation detection in TikTok

Extraction and aggregation of multimodal inherently-interpretable features of video contents to assess disinfodemics [UGR + Andrés Montoro et al. @ UCLM]



- (1) Features are extracted from videos using deep learning
- (2) Embeddings are aggregated using a weighted model (Granular Linguistic Model of Phenomena + Hierarchical Perception Mapping)
- (3) A fuzzy "suspiciousness" score is calculated, applicable to disinformation & conspiracies
- (4) An LLM is used to generate reports in natural language

Tackling COVID-19 **conspiracy** on Twitter

- Shared task at MediaEval 2022
- **Twitter** data: scraping, keyword-filtering, cleaning, annotation
- User graph: nodes are users, edges are user-user interactions
- **Text-based detection** of conspiracy theories
- **Graph-based conspiracy spreader detection**
- **Conspiracy categories:** suppressed cures, behaviour and mind control, antivax, fake virus, intentional pandemic, harmful radiation or influence, population reduction, new world order, and satanism
- Text-conspiracy relation: support, mention, no-mention

<https://github.com/konstapo/2022-Fake-News-MediaEval-Task>

Langguth J., Schroeder D.T., Filkuková P., Brenner S., Phillips J., Pogorelov K. (2023).

COCO: An Annotated Twitter Dataset of COVID-19 Conspiracy Theories. Journal of Computational Social Science.

PRHLT at MediaEval 2022 (task 1)

- Given a **tweet and a conspiracy theory** decide if:
 1. There is no mention of the conspiracy in the text
 2. The text mentions the conspiracy but does not support it
 3. The text supports the conspiracy
- Results achieved by the top 4 teams, Matthews Correlation Coefficient (MCC) metric:

Team	MCC Score
Korenčić et al. 2023	0.738
Peskine, Papotti, et al. 2023	<u>0.710</u>
Akbari 2023	0.702
Bocconi et al. 2023	0.596

Korenčić D., Grubišić I., Toselli A.H., Chulvi B., Rosso P. (2023).

Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs.

Working Notes Proc. of the MediaEval 2022 Workshop Bergen, Norway

PRHLT at MediaEval 2022 (task 2)

- An undirected $G=(V,E)$ derived from Twitter data; V =users, E =connection between users: 1,679,011 nodes, 268,694,698 edges, avg. 160 edges/node
- Label users as conspiracy spreaders or non-conspiracy spreaders
- Train set (1,913 users), Test set (830 users)
- Results achieved by the top 4 teams (MCC):

Team	MCC Score
Jiménez et al. 2023	0.434
Peskine, Papotti, et al. 2023	<u>0.355</u>
Korenčić et al. 2023	0.283
Bocconi et al. 2023	0.110

Jiménez A. G., Panizo A. , Torregrosa J., Camacho D. (2023).

Representational Learning for the Detection of COVID related Conspiracy Spreaders in Online Platforms.

Working Notes Proc. of the MediaEval 2022 Workshop Bergen, Norway

Definitions matters for GPT

Approach	MCC	Precision	Recall	F1
Zero-shot	0.398	0.331	0.852	0.440
w/ Example-generated definitions	0.442	0.371	0.831	0.485
w/ Human-written definitions	0.516	0.464	0.823	0.555
CT-BERT ensembling	0.780	0.779	0.849	0.810

Peskine Y., Korencic D., Grubišic I., Papotti P., Troncy R., Rosso P. (2023)

Definitions Matter: Guiding GPT for Multi-label Classification. In: Findings of EMNLP-2023

Taxonomies on conspiracy theories

- **Focus:**

- **outsiders vs insiders** (exogroup vs endogroup) as **friend/enemy schema**
- **Social Identity Theory** that gives to the individual a social identity and a **sense of belonging**

- **Drawbacks:**

- it mixes **actions** and actors, i.e. groups of people (**social categories**): an event (e.g. AIDS) may provoke the *action* of a *social group*
- **actors** with **consequences** and **objectives** (labelled the three of them with just one label: **insiders**)
- mixing **actors and actions** cannot capture an intergroupal conflict, just friend/enemy schema

Holur P., Wang T., Shahsavari S., Tangherlini T., Roychowdhury V. (2022).

Which Side are you On? Insider-Outsider Classification in Conspiracy-theoretic Social Media.

Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 4975 - 4987

Conspiracy narrative vs critical thinking

- **Social psychologist** on board (UPV) and **linguists** (UB) for the annotation
- **"Us vs them"** narrative
- **Insiders** include **campaigners** and **victims**
- **Outsiders** include **agents** and **facilitators**
- Categories at **span level**
- **Domain-agnostic**: it could be applied to other conspiracy theories (e.g. climate change)

Korenčić D., Chulvi B., Bonet X., Taulé M., Toselli A.H., Rosso P. (2024).

What Distinguishes Conspiracy from Critical Narratives? A Computational Analysis of Oppositional Discourse.




Expert Systems (accepted)

Conspiracy narrative vs critical thinking

- **Agents (A)**
- **Objectives (O)**
- **Consequences (CN)**
- **Victims (V)**
- **Campaigners (CM)** : activists
- **Facilitators (F)** : collaborators with conspiracy propagators (conspiracy narrative) vs implementing measures dictated by the authorities (critical thinking)

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 **CN**, 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 **CN** by the media **F** and governments **F**. To those awake already **CM**, 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 **CN**. These are very evil beings **A**, intent on destroying us **O** regular people **V**.

Oppositional thinking analysis: Conspiracy narrative vs critical thinking

- **Telegram:** 5k messages in each language  
- Oppositional non-mainstream views on the **COVID-19 pandemic**
- Shared task at  **PAN** 2024
- 1st task: **conspiracy theories vs critical thinking narratives**
(Matthew's correlation coefficient)
- 2nd task: **text-span recognition of elements of oppositional narratives**
(macro-averaged span-F1)
- **83 teams** participated <https://pan.webis.de/clef24/pan24-web/oppositional-thinking-analysis.html>

Korenčić D., Chulvi B., Bonet X., Taulé M., Rosso P., Rangel F. (2024).

Overview of the Oppositional Thinking Analysis PAN Task at CLEF 2024.

Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum.

Gracias



XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (PLEC2021-007681)

proso@dsic.upv.es